Optimal Aggregation of Prediction Intervals under Unsupervised Domain Shift

Jiawei Ge*
Operations Research & Financial Engineering
Princeton University
jg5300@princeton.edu

Debarghya Mukherjee*
Department of Mathematics and Statistics
Boston University
mdeb@bu.edu

Jianqing Fan
Operations Research & Financial Engineering
Princeton University
jqfan@princeton.edu

Abstract

As machine learning models are increasingly deployed in dynamic environments, it becomes paramount to assess and quantify uncertainties associated with distribution shifts. A distribution shift occurs when the underlying data-generating process changes, leading to a deviation in the model's performance. The prediction interval, which captures the range of likely outcomes for a given prediction, serves as a crucial tool for characterizing uncertainties induced by their underlying distribution. In this paper, we propose methodologies for aggregating prediction intervals to obtain one with minimal width and adequate coverage on the target domain under unsupervised domain shift, under which we have labeled samples from a related source domain and unlabeled covariates from the target domain. Our analysis encompasses scenarios where the source and the target domain are related via i) a bounded density ratio, and ii) a measure-preserving transformation. Our proposed methodologies are computationally efficient and easy to implement. Beyond illustrating the performance of our method through real-world datasets, we also delve into the theoretical details. This includes establishing rigorous theoretical guarantees, coupled with finite sample bounds, regarding the coverage and width of our prediction intervals. Our approach excels in practical applications and is underpinned by a solid theoretical framework, ensuring its reliability and effectiveness across diverse contexts.

1 Introduction

In the modern era of big data and complex machine learning models, extensive data collected from diverse sources are often used to build a predictive model. However, the assumption of independent and identically distributed (i.i.d.) data is frequently violated in practical scenarios. Take algorithmic fairness as an example: historical data often exhibit sampling biases towards certain groups, like females being underrepresented in credit card data. Over time, the differences in group proportions have diminished, leading to distribution shifts. Consequently, models trained on historical data may face shifted distributions during testing, and proper adjustments are needed. Distribution shift has garnered significant attention from statistical and machine learning communities under various names, i.e., transfer learning [PY09, WKW16], domain adaptation [FVRA21], domain generalization [ZLQ+22, WLL+22], continual learning [DLAM+21, MLJ+22], multitask learning [ZY21]

*equal contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2341

etc. While numerous methods are available in the literature for training predictive models under distribution shift, uncertainty quantification under distribution shift has received relatively scant attention despite its crucial importance. One notable exception is conformal prediction under distribution shift; [TFBCR19] proposed a variant of standard conformal inference methods to accommodate test data from a distinct distribution from the training data under the covariate shift. Recently, [GC21] introduced an adaptive conformal inference approach suitable for continuously changing distributions over time. Additionally, quantile regression under distribution shift offers another avenue for addressing uncertainty quantification under distribution shift [ERS+22].

Although few methods exist for constructing prediction intervals under distribution shift, most focus primarily on ensuring coverage guarantee rather than minimizing interval width. This prompts the immediate question:

Can we generate prediction intervals in the target domain that provide both i) coverage guarantee and ii) minimal width?

This paper seeks to address this question by leveraging model aggregation techniques [NW15, MNW16, CEN14, Vov15, HKNC14]. Suppose we have K different methods for constructing prediction intervals in the *source* domain. Our proposed approach efficiently combines these methods to produce prediction intervals in the *target* domain with adequate coverage and minimal width. When individual methods are the elementary basis functions, such as the kernel basis, the resulting aggregation is indeed a construction of the prediction interval based on the basis functions. Our methodology draws inspiration primarily from recent work [FGM23] on prediction interval aggregation under the i.i.d. setting. However, a key distinction lies in our focus on *unsupervised domain adaptation*, where we can access labeled samples from the source and unlabeled samples from the target domain. Certain assumptions regarding the similarities between these domains are necessary to facilitate knowledge transfer from the source to the target domain. We explore two types of similarities in this paper: i) *covariate shift*, where we assume that the distribution of the response variable Y given X is consistent across both domains, albeit the distribution of X may differ, and ii) *domain shift*, where we assume that the conditional distribution of Y given X remains unchanged up to a measure-preserving transformation. Covariate shift is a well-explored concept in transfer learning and has also garnered attention in uncertainty quantification. It allows different distributions of X while maintaining identical conditional distributions $Y|X$ across domains. For constructing conformal prediction intervals within this framework, see [TFBCR19, HL23, YKT22, LC21] and references therein. On the other hand, *distribution shift* is more general, allowing both the distribution of X and the conditional distribution of $Y|X$ to differ across domains. Our methods in this context draw upon domain matching principles via transport map, as proposed in [CFT14] and further elaborated in subsequent works like [CFTR16, CFHR17, RHS17], among others. The key assumption is the existence of a measure-preserving/domain-aligning map T from the target to the source domain, such that the conditional distribution of $Y|X$ on the target domain matches $Y|T(X)$ on the source domain, i.e., conditional distributions matches upon domain alignment. The case where the domain-aligning map is the optimal transport map has received considerable attention in the literature, e.g., see [CFT14, CFTR16, CFHR17, XLW+20]. Empirical evidence supports the efficacy of domain alignment through optimal transport maps across various datasets. For instance, in [XLW+20], a variant of this method is applied for domain adaptation in image recognition tasks, such as recognizing similarities between USPS [Hul94], MNIST [LBBH98], and SVHN digit images [NWC+11], as well as between different types of images in the Office-home dataset [VECP17], including artistic and product images. Additionally, in [CFT14], the authors explore the impact of domain alignment via optimal transport maps on the face recognition problem, where different poses give rise to distinct domains. However, most of these works concentrate on training predictors that perform well on the target domain without any guarantee regarding uncertainty quantification. To our knowledge, this is the first work to propose a method with rigorous theoretical guarantees for constructing prediction intervals on the target domain under the domain-aligning assumption within an unsupervised domain adaptation framework. We now summarize our contributions.

Our Contributions: This paper introduces a novel methodology for aggregating various prediction methods available on the source domain to construct a unified prediction interval on the target domain under both covariate shift and domain shift assumptions. Our approach is simple and easy to implement and requires solving a convex optimization problem, which can even be

simplified to a linear program problem in certain scenarios. We also establish rigorous theoretical guarantees, presenting finite sample concentration bounds to demonstrate that our method achieves adequate coverage with a small width. Furthermore, our methodology extends beyond model aggregation; it can be used to construct efficient prediction intervals from any convex collection of candidate functions. In the paper, we adopt this broader perspective, discussing how the aggregation of prediction intervals emerges as a particular case. Lastly, we validate the effectiveness of our approach by analyzing real-world datasets.

We also want to highlight the differences between our method and a related method proposed in [FGM23]. We deal with unsupervised domain adaptation, i.e., we do not observe any label from the target domain, in contrast to [FGM23], which only deals with i.i.d. data. Hence, significant changes in methodology are required to address the domain shift. Furthermore, as pointed out in Section 3, the shift may cause the optimization problem non-convex, for which we need to introduce a convex surrogate (e.g., the hinge function), leading to additional theoretical challenges.

2 Notations and preliminaries

Notation The covariates of the source and the target domains are denoted by $\mathcal{X}_S$ and $\mathcal{X}_T$, respectively, and $\mathcal{X} := \mathcal{X}_S \cup \mathcal{X}_T$. The space of the label is denoted by $\mathcal{Y}$. We use the notation $\mathbb{E}_S$ (resp. $\mathbb{E}_T$) to denote the expectation with respect to the source (resp. target) distribution. The expectation with respect to sample distribution is denoted by $\mathbb{E}_{n,S}$ and $\mathbb{E}_{n,T}$. We use p_S (resp. p_T) to denote the probability density function of X on the source and the target domain, respectively. Throughout the paper, we use c to denote universal constants, which may vary from line to line.

2.1 Problem formulation

Our setup aligns with the unsupervised domain adaption; we assume to have n_S i.i.d. labeled samples $\{X_{S,i}, Y_{S,i}\}_{i=1}^{n_S} \sim \mathbb{P}_S(X,Y)$ from the source domain, and n_T i.i.d. unlabeled samples $\{X_{T,i}\}_{i=1}^{n_T} \sim \mathbb{P}_T(X)$ from the target domain. Given any $\alpha > 0$, ideally, we want to construct a valid prediction interval with minimal width on the target domain:

$$\min_{u,l} \quad \mathbb{E}_T[u(X) - l(X)], \ \text{ s.t. } \mathbb{P}_T\left(l(X) \le Y \le u(X)\right) \ge 1 - \alpha\,. \tag{2.1}$$

In many practical contexts, the preferred prediction interval takes the form of $m(X) \pm g(X)$, where $m(X)$ is a predictor for Y given X (an estimator of $\mathbb{E}_T[Y \mid X]$), and $g(X)$ gauges the uncertainty of the predictor $m(X)$. The optimizer of (2.1) takes this simplified form when the distribution of $Y - \mathbb{E}_T[Y \mid X]$ is symmetric around 0. Moreover, it offers a straightforward interpretation as the pair (m, g) is a predictor and a function quantifying its uncertainty. Within the framework of this simplified prediction interval, we need to estimate m and g. Estimating the conditional mean function m is relatively easy and has been extensively studied; one may use any suitable parametric/non-parametric method. Upon estimating m, we need to estimate g so that the prediction interval $[m(X) \pm g(X)]$ has both adequate coverage and minimal width. This translates into solving the following optimization problem:

$$\min_{f \in \mathcal{F}} \quad \mathbb{E}_T[f(X)], \ \text{ s.t. } \mathbb{P}_T\left((Y - m(X))^2 > f(X)\right) \le \alpha\,. \tag{2.2}$$

Let f_0 be the solution of the above optimization problem. Then the optimal prediction interval is $[m_0(x) \pm \sqrt{f_0(x)}]$. However, the key challenge here is that we do not observe the response variable Y from the target, and consequently, solving (2.2) becomes infeasible. Hence, we must rely on transferring our knowledge acquired from labeled observations in the source domain, which necessitates making certain assumptions regarding the similarity between the two domains. Depending on the nature of these assumptions regarding domain similarity, our findings are presented in two sections: Section 3 addresses covariate shift under the bounded density ratio assumption, while Section 4 considers a more general distribution assumption under measure-preserving transformations. Furthermore, as will be shown later, this problem, though well-defined, is not easily implementable. Therefore, we propose a surrogate convex optimization problem in this paper and provide its theoretical guarantees.

2.2 Complexity measure

The complexity of the function class $\mathcal{F}$ is usually quantified through the Rademacher complexity, defined as follows.

https://doi.org/10.52202/079017-2341

Definition 2.1 (Rademacher complexity). *Let $\mathcal{F}$ be a function class and $\{X_i\}_{i=1}^n$ be a set of samples drawn i.i.d. from a distribution $\mathcal{D}$. The Rademacher complexity of $\mathcal{F}$ is defined as*

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\epsilon,\mathcal{D}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i)\right], \tag{2.3}$$

where $\{\epsilon_i\}_{i=1}^n$ are i.i.d. Rademacher random variables that equals to ± 1 with probability $1/2$ each.

3 Covariate shift with bounded density ratio

Setup and methodology In this section, we focus on the covariate shift problems, where the marginal densities $p_S(X)$ and $p_T(X)$ of the covariates may vary between the source and target domains, albeit the conditional distribution $Y|X$ remains the same. Denote by $m_0(x) = \mathbb{E}_T[Y|X = x] = \mathbb{E}_S[Y|X = x]$, the conditional mean function. For the ease of the presentation, we assume m_0 is known. If unknown, one may use the labeled source data to estimate it using a suitable parametric/non-parametric estimate (e.g., splines, local polynomial, or deep neural networks), subsequently substituting m_0 with $\hat{m}$ in our approach. The density ratio of the source and the target distribution of X is denoted by $w_0(x) := p_T(x)/p_S(x)$. We henceforth assume that the density ratio is uniformly bounded:

Assumption 3.1. *There exists W such that* $\sup_{x\in\mathcal{X}_S} w_0(x) \le W$.

If w_0 is known, (2.2) has the following sample level counterpart:

$$\min_{f\in\mathcal{F}} \ \mathbb{E}_{n,T}[f(X)], \ \text{ s.t. } \mathbb{E}_{n,S}\left[w_0(X)\mathbb{1}_{(Y-m_0(X))^2>f(X)}\right] \le \alpha\,, \tag{3.1}$$

which is NP-hard owing to the presence of the indicator function. However, in many practical scenarios, it is observed that the shape of the prediction band does not change much if we change the level of coverage (i.e., α); only the bands shrink/expand. Indeed, the true shape determines the average width; if the shape is wrong, then the width of the prediction band is quite likely to be unnecessarily large. Therefore, to obtain a prediction interval with adequate coverage and minimal width, one should first identify the shape of the prediction band and then shrink/expand it appropriately to get the desired coverage. This motivates the following two steps procedure:

Step 1: (Shape estimation) Obtain an initial estimate $\hat{f}_{\text{init}}$ by solving (3.1) for $\alpha = 0$ (to capture the shape):

$$\min_{f\in\mathcal{F}} \ \mathbb{E}_{n,T}[f(X)]\,, \ \text{ s.t. } \ f(X_i) \ge (Y_i - m_0(X_i))^2 \ \ \forall\, 1 \le i \le n_S : w_0(X_i) > 0\,. \tag{3.2}$$

Step 2: (Shrinkage) Refine $\hat{f}_{\text{init}}$ by scaling it down using $\hat{\lambda}(\alpha)$, defined as:

$$\hat{\lambda}(\alpha) = \inf\left\{\lambda \ge 0 : \mathbb{E}_{n,S}[w_0(X)\mathbb{1}_{(Y-m_0(X))^2>\lambda\hat{f}_{\text{init}}(X)}] \le \alpha\right\}. \tag{3.3}$$

The final prediction interval is:

$$\widehat{\text{PI}}_{1-\alpha}(x) = \left[m_0(x) - \sqrt{\hat{\lambda}(\alpha)\hat{f}_{\text{init}}(x)}, m_0(x) + \sqrt{\hat{\lambda}(\alpha)\hat{f}_{\text{init}}(x)}\right]. \tag{3.4}$$

In Step 1, we relax (3.1) by effectively setting $\alpha = 0$. This relaxation aids in determining the optimal shape while also converting (3.1) into a convex optimization problem (equation (3.2)) as long as $\mathcal{F}$ is a convex collection of functions. Furthermore, in (3.2), we only consider those source observations for which $w_0(x) > 0$, as otherwise, the samples are not informative for the target domain. In practice, w_0 is typically unknown; one may use the source and target domain covariates to estimate w_0. Various techniques are available for estimating the density ratio (e.g., [USS+16, CMSE22, Qin98, GSH+08] and references therein). However, any such estimator $\hat{w}(x)$ can be non-zero for x where $w_0(x) = 0$ due to estimation error. Consequently, $\hat{w}$ may not be efficient in selecting informative source samples. To mitigate this issue, we propose below a modification of (3.2), utilizing a hinge function $h_\delta(t) := \max\{0, (t/\delta) + 1\}$:

$$\begin{aligned} &\min_{f\in\mathcal{F}} \ \mathbb{E}_{n,T}[f(X)] \\ &\text{subject to } \ \mathbb{E}_{n,S}[\hat{w}(X)h_\delta\left((Y - m_0(X))^2 - f(X)\right)] \le \epsilon, \end{aligned} \tag{3.5}$$

with δ and ϵ should be chosen based on sample size n_S and the estimation accuracy of $\hat{w}$. When $\hat{w} = w_0$ (i.e., the density ratio is known), then by choosing $\epsilon = 0$ and $\delta \to 0$, (3.5) recovers (3.2). As h_δ is convex, the optimization problem (3.5) is still a convex optimization problem. We summarize our algorithm in Algorithm 1.

Algorithm 1 Prediction intervals with bounded density ratio

1: **Input:** m_0 (or $\hat{m}$ if unknown), density ratio estimator $\hat{w}$, function class $\mathcal{F}$, sample $\mathcal{D}_S = \{(X_{S,i}, Y_{S,i})\}_{i=1}^{n_S}$ and $\mathcal{D}_T = \{X_{T,i}\}_{i=1}^{n_T}$, parameters δ, ϵ, coverage level $1-\alpha$.
2: Obtain $\hat{f}_{\text{init}}$ by solving (3.5).
3: Obtain the shrink level $\hat{\lambda}(\alpha)$ by solving (3.3) with w_0 replaced by $\hat{w}$.
4: **Output:** $\widehat{\text{PI}}_{1-\alpha}(x)$ defined in (3.4).

Theoretical results We next present theoretical guarantees of the prediction interval obtained via Algorithm 1. For technical convenience, we resort to data-splitting; we divide the source data into two equal parts ($\mathcal{D}_{S,1}$ and $\mathcal{D}_{S,2}$), use $\mathcal{D}_{S,1}$ and $\mathcal{D}_T$ to solve (3.5), and $\mathcal{D}_{S,2}$ to obtain the shrink level $\hat{\lambda}(\alpha)$. Without loss of generality, we assume $m_0 \equiv 0$ (otherwise, we set $Y \leftarrow Y - m_0(X)$). A careful inspection of Step 1 reveals that $\hat{f}_{\text{init}}$ aims to approximate a function f^* defined as follows:

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_T[f(X)] \text{ subject to } Y^2 < f(X) \text{ almost surely on target domain}. \tag{3.6}$$

In other words, $\hat{f}_{\text{init}}$ estimates f^* that has minimal width among all functions covering the response variable. This is motivated by the philosophy that the *right shape leads to a smaller width*. The following theorem provides a finite sample concentration bound on the approximation error of $\hat{f}_{\text{init}}$:

Theorem 3.2. *Suppose $Y^2 - f^*(X) \le B$ on the source domain and has a density bounded by L. Also assume $\|f\|_\infty \le B_{\mathcal{F}}$ for all $f \in \mathcal{F}$. Then for*

$$\epsilon \ge L\delta + W\sqrt{\tfrac{t}{n_S}} + \tfrac{B+\delta}{\delta} \cdot \left(\mathbb{E}_S\left[|\hat{w}(X) - w_0(X)|\right] + (W + W')\sqrt{\tfrac{t}{n_S}} \right), \tag{3.7}$$

we have with probability at least $1 - 3e^{-t}$:

$$\mathbb{E}_T[\hat{f}_{\text{init}}(X)] \le \mathbb{E}_T[f^*(X)] + 2\mathcal{R}_{n_T}(\mathcal{F} - f^*) + 2B_{\mathcal{F}}\sqrt{\tfrac{t}{2n_T}}$$

where $W' = \|\hat{w}\|_\infty$.

The bound in the above theorem depends on the Rademacher complexity of $\mathcal{F}$ (the smaller, the better), the estimation error of w_0, and an interplay between the choice of (ϵ, δ). The lower bound on ϵ in (3.7) depends on both δ and $1/\delta$. Although it is not immediate from the above theorem why we need to choose ϵ to be as small as possible, it will be apparent in our subsequent analysis; indeed if ϵ is large in (3.5), then $\hat{f}_{\text{init}} \equiv 0$ will be a solution of (3.5). Consequently, the shape will not be captured. Therefore, one should first choose δ (say δ^*), that minimizes the lower bound (3.7), and then set $\epsilon = \epsilon^*$ equal to the value of the right-hand side of (3.7) with $\delta = \delta^*$, which ensures that ϵ^* is optimally defined to capture the shape accurately. Once the shape is identified, we shrink it properly in Step 2 to attain the desired coverage and reduce the width. Although ideally $\hat{\lambda}(\alpha) \le 1$, it is not immediately guaranteed as we use separate data ($\mathcal{D}_{S,2}$) for shrinking. The following lemma shows that $\hat{\lambda}(\alpha) \le 1$ for any fixed $\alpha > 0$ as long as the sample size is large enough. Recall that the data were split into exactly half with size $n_S = |\mathcal{D}_S|$.

Lemma 3.3. *Under the aforementioned choice of (ϵ^*, δ^*), we have with high probability:*

$$\frac{1}{n_S/2} \sum_{i \in \mathcal{D}_{S,2}} \hat{w}(X_i) \mathbb{1}_{\{(Y_i - m_0(X_i))^2 > \hat{f}_{\text{init}}(X_i)\}} \le \alpha,$$

for all large n_S, provided that $\hat{w}$ is a consistent estimator of w_0. Hence, $\hat{\lambda}(\alpha) \le 1$.

Our final theorem for this section provides a coverage guarantee for the prediction interval given by Algorithm 1.

Theorem 3.4. *For the prediction interval obtained in (3.4), with probability greater than $1 - 2e^{-t}$:*

$$\left| \mathbb{P}_T\left(Y^2 > \hat{\lambda}(\alpha)\hat{f}_{\text{init}}(X) \mid \mathcal{D}_S \cup \mathcal{D}_T \right) - \alpha \right| \le \mathbb{E}_S\left[|\hat{w}(X) - w(X)|\right] + (2W + W')\sqrt{\tfrac{t}{2n_S}} + \sqrt{\tfrac{C}{n_S}}$$

for some constant $C > 0$ and $W' = \|\hat{w}\|_\infty$.

https://doi.org/10.52202/079017-2341

Theorem 3.4 validates the coverage of the prediction interval derived through Algorithm 1, achieving the desired coverage level as the estimate of w_0 improves and sample size expands. Theorems 3.2 and 3.4 collectively demonstrate the efficacy of our method in maintaining validity and accurately capturing the optimal shape of the prediction band, which in turn leads to small interval widths.

Remark 3.5. *In our optimization problem, we've substituted the indicator loss with the hinge loss function to ensure convexity. However, it's worth noting that if we know the subset of $\mathcal{X}_S$ where $w_0(x) > 0$ beforehand, we could directly optimize (3.2). This approach would be easy to implement and wouldn't involve tuning parameters (δ, ϵ). A special case is when $w_0(x) > 0$ for all $x \in \mathcal{X}_S$ (as is true in our experiment), which simplifies the condition in (3.2) to $f(X_i) \geq (Y_i - m_0(X_i))^2$ for all $1 \leq i \leq n_S$. However, if this information is unavailable, one can still employ (3.2) by enforcing the constraint on all source observations. While this approach might result in wider prediction intervals, it is easy to implement and doesn't require tuning parameters.*

4 Domain shift and transport map

Setup and methodology In the previous section, we assume a uniform bound on the density ratio. However, this may not be the case in reality; it is possible that there exists $x \in \mathsf{supp}(\mathcal{X}_T) \cap \mathsf{supp}(\mathcal{X}_S^c)$, which immediately implies that $w_0(x) = \infty$. In image recognition problems, if the source data are images taken during the day at some place, and the target data are images taken at night, then this directly results in an unbounded density ratio (due to the change in the background color). Yet a transport map could effectively model this shift by adapting features from the source to correspond with those of the target, maintaining the underlying patterns or object recognition capabilities across both domains. To perform transfer learning in this setup, we model the domain shift via a measure transport map T_0 that preserves the conditional distribution, as elaborated in the following assumption:

Assumption 4.1. *There exists a measure transport map $T_0 : \mathcal{X}_T \to \mathcal{X}_S$, i.e., $T_0(X_T) \overset{d}{=} X_S$, such that: $\mathbb{P}_T(Y \mid X = x) \overset{d}{=} \mathbb{P}_S(Y \mid X = T_0(x))$, $\forall x \in \mathcal{X}_T$.*

This assumption allows the extrapolation of source domain information to the target domain via T_0, enabling the construction of prediction intervals at $x \in \mathcal{X}_T$ by leveraging the analogous intervals at $T_0(x) \in \mathcal{X}_S$. Inspired by this observation, we present our methodology in Algorithm 2 that essentially consists of two key steps: i) constructing a prediction interval in the source domain and ii) transporting this interval to the target domain using the estimated transport map T_0. If T_0 (or its estimate) is not given, it must be estimated from the source and the target covariates. Various methods are available in the literature (e.g., [DNWP22, SDF+17, MTOL20, DGS21]), and practitioners can pick a method at their convenience. Notably, the processes described in equations (4.1) and (4.2) follow the methodology (i.e., (3.2) and (3.3)) from Section 3 for scenarios without shift (i.e., $w_0 \equiv 1$), adding a slight δ to ensure coverage even when $\mathcal{F}$ is complex. In Algorithm 2, we assume

Algorithm 2 Transport map

1: **Input:** conditional mean function m_0 on the source domain, transport map estimator $\hat{T}_0$, function class $\mathcal{F}$, sample $\mathcal{D}_S = \{(X_{S,i}, Y_{S,i})\}_{i=1}^{n_S}$ and $\mathcal{D}_T = \{X_{T,i}\}_{i=1}^{n_T}$, parameter δ, coverage level $1 - \alpha$.

2: Obtain $\hat{f}_{\text{init}}$ by solving:

$$\min_{f \in \mathcal{F}} \ \frac{1}{n_S} \sum_{i=1}^{n_S} f(X_{S,i}) \,, \ \text{ s.t. } \ f(X_{S,i}) \geq (Y_{S,i} - m_0(X_{S,i}))^2 \ \forall\, i \in [n_S] \,. \tag{4.1}$$

3: Obtain the shrink level

$$\hat{\lambda}(\alpha) := \inf \left\{ \lambda > 0 : \frac{1}{n_S} \sum_{i=1}^{n_S} \mathbb{1}_{(Y_{S,i} - m_0(X_{S,i}))^2 \geq \lambda(\hat{f}_{\text{init}}(X_{S,i}) + \delta)} \leq \alpha \right\} . \tag{4.2}$$

4: **Output:** $\widehat{\mathrm{PI}}_{1-\alpha}(x) = \left[m_0 \circ \hat{T}_0(x) \pm \sqrt{\hat{\lambda}(\alpha) \cdot \left(\hat{f}_{\text{init}} \circ \hat{T}_0(x) + \delta \right)} \right]$.

the conditional mean function m_0 on the source domain is known. In cases where the conditional

mean function m_0 on the source domain is unknown, it can be estimated using standard regression methods from labeled source data, after which m_0 is replaced by this estimate, $\hat{m}$.

Remark 4.2 (Model aggregation)**.** *Suppose we have K different methods $\{f_1, \ldots, f_K\}$ for constructing prediction intervals in the source domain. In the context of model aggregation, (4.1) then reduces to:*

$$\begin{aligned} \min_{\alpha_1,\ldots,\alpha_K} \quad & \frac{1}{n_S}\sum_{i=1}^{n_S}\Big\{\sum_{j=1}^{K}\alpha_j f_j(X_{S,i})\Big\} \\ \text{subject to} \quad & \sum_{j=1}^{K}\alpha_j f_j(X_{S,i}) \geq (Y_{S,i} - m_0(X_{S,i}))^2 \;\forall\, i \in [n_S]\,, \\ & \alpha_j \geq 0, \quad \forall\, 1 \leq j \leq K\,. \end{aligned}$$

In other words, the function class $\mathcal{F}$ is a linear combination of the candidate methods. The problem is then simplified to a linear program problem, which can be implemented efficiently using standard solvers.

Theoretical results We now present theoretical guarantees of our methodology to ensure that our method delivers what it promises: a prediction interval with adequate coverage and small width. For technical simplicity, we split data here: divide the labeled source observation with two equal parts (with $n_S/2$ observations in each), namely $\mathcal{D}_{S,1}$ and $\mathcal{D}_{S,2}$. We use $\mathcal{D}_{S,1}$ to solve (4.1) and obtain the initial estimator $\hat{f}_{\text{init}}$, and $\mathcal{D}_{S,2}$ to solve (4.2), i.e. obtaining the shrinkage factor $\hat{\lambda}(\alpha)$. Henceforth, without loss of generality, we assume $m_0 = 0$ and present the theoretical guarantees of our estimator. We start with an analog of Theorem 3.2, which ensures that with high probability $\hat{f}_{\text{init}} \circ \hat{T}_0$ approximates the function that has minimal width among all the functions in $\mathcal{F}$ composed with T_0 that covers the labels on the target almost surely:

Theorem 4.3. *Assume the function class $\mathcal{F}$ is $B_{\mathcal{F}}$-bounded and $L_{\mathcal{F}}$-Lipschitz. Define*

$$\Delta = \min\left\{\mathbb{E}_T[f \circ T_0(X)] : f \in \mathcal{F}, Y^2 \leq f \circ T_0(X) \text{ a.s. on target domain}\right\}\,.$$

Then we have with probability $\geq 1 - e^{-t}$:

$$\mathbb{E}_T[\hat{f}_{\text{init}} \circ \hat{T}_0(X)] \leq \Delta + 4\mathcal{R}_{n_S}(\mathcal{F}) + L_{\mathcal{F}}\mathbb{E}_T[|\hat{T}_0(X) - T_0(X)|] + 4B_{\mathcal{F}}\sqrt{\frac{t}{2n_S}}\,.$$

The upper bound on the population width of $\hat{f}_{\text{init}} \circ \hat{T}_0(x)$ consists of four terms: the first term is the *minimal possible width* that can be achieved using the functions from $\mathcal{F}$, the second term involves the Rademacher complexity of $\mathcal{F}$, the third term encodes the estimation error of T_0, and the last term is the deviation term that influences the probability. Hence, the margin between the width of the predicted interval and the minimum achievable width is small, with the convergence rate relying on the precision of estimating T_0 and the complexity of $\mathcal{F}$, as expected.

We next establish the coverage guarantee of our estimator of Algorithm 2, obtained upon suitable truncation of $\hat{f}_{\text{init}}$. As mentioned, the shrinkage operation is performed on a separate dataset $\mathcal{D}_{S,2}$. Therefore, it is not immediate whether the shrinkage factor $\hat{\lambda}(\alpha)$ is smaller than 1, i.e., whether we are indeed shrinking the confidence interval ($\hat{\lambda}(\alpha) > 1$ is undesirable, as it will widen $\hat{f}_{\text{init}}$, increasing the width of the prediction band). The following lemma shows that with high probability, $\hat{\lambda}(\alpha) \leq 1$.

Lemma 4.4. *With probability greater than or equal to $1 - e^{-t}$, we have:*

$$\mathbb{P}(\hat{\lambda}(\alpha) > 1 \mid \mathcal{D}_{S,1}, \mathcal{D}_T) \leq e^{-\frac{(\alpha - p_{n_S})^2 n_S}{6p_{n_S}}}\,,$$

where

$$p_{n_S} = \mathbb{P}_S\left(Y^2 \geq \hat{f}_{\text{init}}(X) + \delta \,\middle|\, \mathcal{D}_{S,1}, \mathcal{D}_T\right) \leq \tfrac{4}{\delta}\left(\sqrt{\tfrac{\mathbb{E}_S[Y^4]}{n_S}} + \mathcal{R}_{n_S}(\mathcal{F})\right) + \sqrt{\tfrac{t}{n_S}}\,.$$

https://doi.org/10.52202/079017-2341

Here p_{n_S} is the conditional probability of a test observation Y falling outside $[-\sqrt{\hat{f}_{\text{init}}(X)+\delta}, \sqrt{\hat{f}_{\text{init}}(X)+\delta}]$, which is small as evident from the above lemma. In particular, for model aggregation, if $\mathcal{F}$ is the linear combination of K functions, then p_{n_S} is of the order $\sqrt{K/n_S}$. Hence, the final prediction interval is guaranteed to be a compressed form of $\hat{f}_{\text{init}}$ with an overwhelmingly high probability. We present our last theorem of this section, confirming that the prediction interval derived from Algorithm 2 achieves the intended coverage level with a high probability:

Theorem 4.5. *Under the same setup of Theorem 4.3, along with the assumption that $f_S(y \mid x)$ is uniformly bounded by G, we have with probability greater than $1 - cn_S^{-10}$ that*

$$\left|\mathbb{P}_T\left(Y^2 \geq \hat{\lambda}(\alpha)\left(\hat{f}_{\text{init}} \circ \hat{T}_0(X) + \delta\right) \mid \mathcal{D}_S \cup \mathcal{D}_T\right) - \alpha\right| \leq C\sqrt{\frac{\log n_S}{n_S}} + GL_{\mathcal{F}} \cdot \mathbb{E}_T\left[\left|\hat{T}_0(X) - T_0(X)\right|\right].$$

As for Theorem 4.3, the bound obtained in Theorem 4.5 also depends on two crucial terms: Rademacher complexity of $\mathcal{F}$ and estimation error of T_0. Therefore, the key takeaway of our theoretical analysis is that the prediction interval obtained from Algorithm 2 asymptotically achieves nominal coverage guarantee and minimal width. Furthermore, the approximation error intrinsically depends on the Rademacher complexity of the underlying function class and the precision in estimating T_0.

Remark 4.6 (Measure preserving transformation)**.** *In our approach, T_0 is employed to maintain measure transformation, although it may not necessarily be an optimal transport map. Yet, estimating T_0 can be challenging in many practical scenarios. In such cases, simpler transformations like linear or quadratic adjustments are often utilized to align the first few moments of the distributions. Various methods provide such simple solutions, including, but not limited to, CORAL [SFS17] and ADDA [THSD17].*

5 Application

In this section, we illustrate the effectiveness of our method by applying it to five different datasets: i) airfoil dataset [DG19], ii) real estate data [Yeh18], iii) energy efficiency data [TX12b], iv) appliance energy prediction data [Can17], and v) ET Dataset (ETT-small) [ZZP+21]. The first four datasets are freely available in the UCI repository, and the last dataset can be found in this GitHub link. Here, we illustrate the procedure using the airfoil dataset, and the details of our experiments using the other datasets can be found in Appendix C. The airfoil dataset includes 1503 observations, featuring a response variable Y (scaled sound pressure level) and a five-dimensional covariate X (log of frequency, angle of attack, chord length, free-stream velocity, log of suction side displacement thickness). We assess and compare the performance of our prediction intervals in terms of coverage and width with those generated by the weighted split conformal prediction method described in [TFBCR19]. We use the same data-generating process described in [TFBCR19] to facilitate a direct comparison. We run experiments 200 times; each time, we randomly partition the data into two parts $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, where $\mathcal{D}_{\text{train}}$ contains 75% of the data, and $\mathcal{D}_{\text{test}}$ contains 25% of the data. Following [TFBCR19], we *shift* the distribution of the covariates of $\mathcal{D}_{\text{test}}$ by weighted sampling with replacement, where the weights are proportional to

$$w(x) = \exp(x^T\beta), \quad \text{where} \quad \beta = (-1, 0, 0, 0, 1).$$

These reweighted observations in $\mathcal{D}_{\text{test}}$, which we call $\mathcal{D}_{\text{shift}}$, act as observations from the target domain. Clearly, by our data generation mechanism $w_0(x) = f_T(x)/f_S(x) = c\exp(x^\top\beta)$, where c is the normalizing constant. The source and target domains share the same support under this configuration. As our methodology is developed for unsupervised domain adaptation, we do not use the label information of $\mathcal{D}_{\text{shift}}$ to develop the target domain's prediction interval.

Density ratio estimation We use the probabilistic classification technique to estimate the density based on the source and the target covariates. Let $X_1, \ldots, X_{n_1}$ be the covariates in dataset $\mathcal{D}_{\text{train}}$ and $X_{n_1+1}, \ldots, X_{n_1+n_2}$ be the covariates in dataset D_{shift}. The density ratio estimation proceeds in two steps: (1) logistic regression is applied to the feature-class pairs $\{(X_i, C_i)\}_{i=1}^n$, where $C_i = 0$ for

$i = 1, \ldots, n_1$ and $C_i = 1$ for $i = n_1 + 1, \ldots, n_1 + n_2$, yielding an estimate of $\mathbb{P}(C = 1 \mid X = x)$, denoted as $\hat{p}(x)$; (2) the density ratio estimator is then defined as $\hat{w}(x) = \frac{n_1}{n_2} \cdot \frac{\hat{p}(x)}{1-\hat{p}(x)}$. Further explanations are provided in Appendix B.

Implementation of our method and results As the mean function $m_0(x) = \mathbb{E}[Y \mid X = x]$ (which is the same on the source and the target domain) is unknown, we first estimate it via linear regression, which henceforth will be denoted by $\hat{m}(x)$. To construct a prediction interval, we consider the model aggregation approach, i.e., the function class $\mathcal{F}$ is defined as the linear combination of the following six estimates:

(1) **Estimator 1**(f_1): A neural network based estimator with depth=1, width=10 that estimates the 0.85 quantile function of $(Y - \hat{m}(X))^2 \mid X = x$.

(2) **Estimator 2**(f_2): A fully connected feed forward neural network with depth=2 and width=50 that estimates the 0.95 quantile function of $(Y - \hat{m}(X))^2 \mid X = x$.

(3) **Estimator 3**(f_3): A quantile regression forest estimating the 0.9 quantile function of $(Y - \hat{m}(X))^2 \mid X = x$.

(4) **Estimator 4**(f_4): A gradient boosting model estimating the 0.9 quantile function of $(Y - \hat{m}(X))^2 \mid X = x$.

(5) **Estimator 5**(f_5): An estimate of $\mathbb{E}[(Y - \hat{m}(X))^2 \mid X = x]$ using random forest.

(6) **Estimator 6**(f_6): The constant function 1.

Here, the quantile estimators are obtained by minimizing the corresponding check loss. The implementation of our method is summarized as follows: (1) We divide the training data $\mathcal{D}_{\text{train}}$ into two halves $\mathcal{D}_1 \cup \mathcal{D}_2$. We utilize dataset $\mathcal{D}_1$ to derive a mean estimator and six aforementioned estimates. We also employ the covariates from $\mathcal{D}_1$ and D_{shift} to compute a density ratio estimator. (2) We further split $\mathcal{D}_2$ into two equal parts $\mathcal{D}_{2,1}$ and $\mathcal{D}_{2,2}$. $\mathcal{D}_{2,1}$, along with covariates from $\mathcal{D}_{\text{shift}}$, is used to find the optimal aggregation of the six estimates to capture the shape, i.e., for obtaining $\hat{f}_{\text{init}}$. The second part $\mathcal{D}_{2,2}$ is used to shrink the interval to achieve $1 - \alpha = 0.95$ coverage, i.e. to estimate $\hat{\lambda}(\alpha)$. (3) We evaluate the effectiveness of our approach in terms of the coverage and average bandwidth on the D_{shift} dataset.

In Figure 1, we present the histograms of the coverage and the average bandwidth of our method, and a more general version of weighted conformal prediction in [TFBCR19] over 200 experiments (see Appendix B for details), which show that our method consistently yields a shorter prediction interval than the weighted conformal prediction while maintaining coverage. Over 200 experiments, the average coverage achieved by our method was 0.964029 (SD = 0.04), while the weighted conformal prediction method achieved an average coverage of 0.9535 (SD = 0.036). Additionally, the average width of the prediction intervals for our method was 13.654 (SD = 2.22), compared to 20.53 (SD = 4.13) for the weighted conformal prediction. Regarding the performance of intervals over 95% coverage, our method achieved this in 72.5% of cases with an average width of 14.35 (SD = 2.22). In contrast, the weighted conformal prediction method did so in 57% of cases with an average width of 21.4 (SD = 4.39). Boxplots are presented in Appendix B for further comparison.

5.1 Robustness of our method

One of the strengths of our approach lies in its resilience to the misspecification of certain components. The core idea behind our method is to combine multiple predictors to create a prediction interval that ensures sufficient coverage while maintaining a narrow average width. These individual prediction intervals may be based on estimators of conditional quantiles, means, variances, and other metrics. If some of the component prediction intervals exhibit poor performance, whether due to inadequate coverage or excessive width, our method typically assigns them lower weights, making it robust to their deficiencies. In contrast, other conformal methods that heavily depend on a single component tend to underperform, particularly with respect to average width. To illustrate this phenomenon, we evaluate our method under model misspecification using a simple one-dimensional simulation setup.

$$X \sim \mathsf{unif}([-1,1]), \;\; \xi \sim \mathsf{unif}([-1,1]), \;\; X \perp\!\!\!\perp \xi$$
$$Y = \sqrt{1 + 25X^4}\xi\,.$$

https://doi.org/10.52202/079017-2341

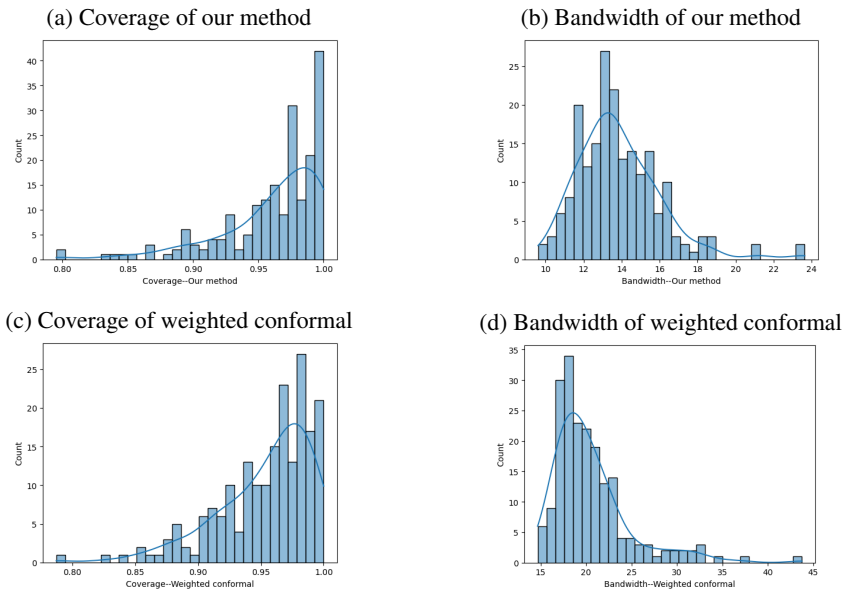


Figure 1: Experiments on Airfoil data using Algorithm 1

Max depth	Avg. width–Our Method	Avg. width–WVAC
3	2.07(0.975)	3.08(0.9712)
5	2.07(0.95)	3.28(0.9664)
7	2.068(0.94)	3.33(0.97)
15	2.08(0.97)	5.00(0.97)

Table 1: Robustness of our method and WVAC. The number inside the parenthesis is the median of coverage over these Monte Carlo iterations.

As mentioned in the previous subsection, our method aggregates six predictor intervals, including an estimator for the conditional variance function (Estimator 5). The weighted variance-adjusted conformal prediction interval (WVAC) relies on accurately estimating this conditional variance. We estimate the conditional variance using a random forest with varying depths ($\{3, 5, 7, 15\}$). For simulation purpose, we generate $n = 2500$ samples, keeping 75% as source data and resampling the remaining 25% with weighted samples proportional to $w(x) \propto (1 + \exp(-2x))^{-1}$. As depth increases, overfitting leads to poor out-of-sample variance predictions. Table 1 summarizes our findings over 100 Monte Carlo iterations, showing that WVAC's average width increases with depth, while our method's average width remains stable.

6 Conclusion

This paper focuses on unsupervised domain shift problems, where we have labeled samples from the source domain and unlabeled samples from the target domain. We introduce methodologies for constructing prediction intervals on the target domain that are designed to ensure adequate coverage while minimizing width. Our analysis includes scenarios in which the source and target domains are related either through a bounded density ratio or a measure-preserving transformation. Our proposed methodologies are computationally efficient and easy to implement. We further establish rigorous finite sample theoretical guarantees regarding the coverage and width of our prediction intervals. Finally, we demonstrate the practical effectiveness of our methodology through its application to the airfoil dataset.

References

[Can17] Luis Candanedo. Appliances Energy Prediction. UCI Machine Learning Repository, 2017. DOI: https://doi.org/10.24432/C5VC8G.

[CEN14] Lars Carlsson, Martin Eklund, and Ulf Norinder. Aggregated conformal prediction. In *Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings 10*, pages 231–240. Springer, 2014.

[CFHR17] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30, 2017.

[CFT14] Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*, pages 274–289. Springer, 2014.

[CFTR16] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

[CMSE22] Kristy Choi, Chenlin Meng, Yang Song, and Stefano Ermon. Density ratio estimation via infinitesimal classification. In *International Conference on Artificial Intelligence and Statistics*, pages 2552–2573. PMLR, 2022.

[DG19] Dheeru Dua and Casey Graff. Uci machine learning repository. https://archive.ics.uci.edu, 2019.

[DGS21] Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Advances in Neural Information Processing Systems*, 34:29736–29753, 2021.

[DLAM+21] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

[DNWP22] Vincent Divol, Jonathan Niles-Weed, and Aram-Alexandre Pooladian. Optimal transport map estimation in general function spaces. *arXiv preprint arXiv:2212.03722*, 2022.

[ERS+22] Cian Eastwood, Alexander Robey, Shashank Singh, Julius Von Kügelgen, Hamed Hassani, George J Pappas, and Bernhard Schölkopf. Probable domain generalization via quantile risk minimization. *Advances in Neural Information Processing Systems*, 35:17340–17358, 2022.

[FGM23] Jianqing Fan, Jiawei Ge, and Debarghya Mukherjee. Utopia: Universally trainable optimal prediction intervals aggregation. *arXiv preprint arXiv:2306.16549*, 2023.

[FVRA21] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pages 877–894, 2021.

[GC21] Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672, 2021.

[GSH+08] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. 2008.

[HKNC14] Mohammad Anwar Hosen, Abbas Khosravi, Saeid Nahavandi, and Douglas Creighton. Improving the quality of prediction intervals through optimal aggregation. *IEEE Transactions on Industrial Electronics*, 62(7):4420–4429, 2014.

 https://doi.org/10.52202/079017-2341

[HL23] Xiaoyu Hu and Jing Lei. A two-sample conditional distribution test using conformal prediction and weighted rank sum. *Journal of the American Statistical Association*, pages 1–19, 2023.

[Hul94] Jonathan J. Hull. A database for handwritten text recognition research. *IEEE Transactions on pattern analysis and machine intelligence*, 16(5):550–554, 1994.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LC21] Lihua Lei and Emmanuel J Candès. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):911–938, 2021.

[LGR+18] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[Mau16] Andreas Maurer. A vector-contraction inequality for rademacher complexities. In *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*, pages 3–17. Springer, 2016.

[MLJ+22] Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469:28–51, 2022.

[MNW16] Katarzyna Maciejowska, Jakub Nowotarski, and Rafał Weron. Probabilistic forecasting of electricity spot prices using factor quantile regression averaging. *International Journal of Forecasting*, 32(3):957–965, 2016.

[MTOL20] Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.

[NW15] Jakub Nowotarski and Rafał Weron. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. *Computational Statistics*, 30:791–803, 2015.

[NWC+11] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 7. Granada, Spain, 2011.

[PY09] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[Qin98] Jing Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630, 1998.

[RHS17] Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings, Part II 10*, pages 737–753. Springer, 2017.

[RPC19] Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[SDF+17] Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

[SFS17] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. *Domain adaptation in computer vision applications*, pages 153–171, 2017.

[TFBCR19] Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

[THSD17] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[TX12a] Athanasios Tsanas and Angeliki Xifara. Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. *Energy and buildings*, 49:560–567, 2012.

[TX12b] Athanasios Tsanas and Angeliki Xifara. Energy Efficiency. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C51307.

[USS+16] Masatoshi Uehara, Issei Sato, Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Generative adversarial nets from a density ratio estimation perspective. *arXiv preprint arXiv:1610.02920*, 2016.

[VECP17] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[Vov15] Vladimir Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74:9–28, 2015.

[WKW16] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3:1–40, 2016.

[WLL+22] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and S Yu Philip. Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072, 2022.

[XLW+20] Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, and Jindong Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4394–4403, 2020.

[Yeh18] I-Cheng Yeh. Real Estate Valuation. UCI Machine Learning Repository, 2018. DOI: https://doi.org/10.24432/C5J30W.

[YH18] I-Cheng Yeh and Tzu-Kuang Hsu. Building real estate valuation models with comparative approach through case-based reasoning. *Applied Soft Computing*, 65:260–271, 2018.

[YKT22] Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift. *arXiv preprint arXiv:2203.01761*, 2022.

[ZLQ+22] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415, 2022.

[ZY21] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2021.

[ZZP+21] H Zhou, S Zhang, J Peng, S Zhang, J Li, H Xiong, and W Zhang Informer. Beyond efficient transformer for long sequence time-series forecasting., 2021, 35. *DOI: https://doi. org/10.1609/aaai. v35i12*, 17325:11106–11115, 2021.

 https://doi.org/10.52202/079017-2341

A Proofs

A.1 Proof of Theorem 3.2

First, we show that for our choice of (ϵ, δ), as depicted in Theorem 3.2, f^* is a feasible solution of equation (3.5). Consider w_0 instead of $\hat{w}$. By definition of f^*,

$$\begin{aligned}\mathbb{P}_T(Y^2 \le f^*(X)) = 1 &\iff \mathbb{E}_S\left[w_0(X)\mathbb{1}_{Y^2 > f^*(X)}\right] = 0 \\ &\iff w_0(X)\mathbb{1}_{Y^2>f^*(X)} = 0 \quad \text{a.s. on the source domain}\,.\end{aligned}$$

This implies:

$$\begin{aligned}&\frac{1}{n_S/2}\sum_{i\in\mathcal{D}_{S,1}} w_0(X_i)h_\delta\left(Y_i^2 - f^\star(X_i)\right) \\ &= \frac{1}{n_S/2}\sum_{i\in\mathcal{D}_{S,1}} w_0(X_i)h_\delta\left(Y_i^2 - f^\star(X_i)\right)\mathbf{1}_{Y_i^2\le f^\star(X_i)} \\ &= \frac{1}{n_S/2}\sum_{i\in\mathcal{D}_{S,1}} w_0(X_i)h_\delta\left(Y_i^2 - f^\star(X_i)\right)\mathbf{1}_{f^\star(X_i)-\delta\le Y_i^2\le f^\star(X_i)} \\ &\le \frac{1}{n_S/2}\sum_{i\in\mathcal{D}_{S,1}} w_0(X_i)\mathbf{1}_{f^\star(X_i)-\delta\le Y_i^2\le f^\star(X_i)},\end{aligned}$$

where the first equality follows from the fact that $w_0(X)\mathbf{1}_{Y^2>f^\star(X)} = 0$ a.s. on the source domain, the second equality follows from the fact that $h_\delta(t)\mathbf{1}_{t<-\delta} = 0$ for all t, and the last inequality follows from the fact that $h_\delta(Y_i^2 - f^\star(X_i)) \le 1$ when $Y_i^2 - f^\star(X_i) \le 0$. Since $w_0(X)\mathbf{1}_{f^\star(X)-\delta\le Y^2\le f^\star(X)} \le W$, by Hoeffding's inequality, we have with probability at least $1 - e^{-t}$:

$$\begin{aligned}\frac{1}{n_S/2}\sum_{i\in\mathcal{D}_{S,1}} w_0(X_i)h_\delta\left(Y_i^2 - f^\star(X_i)\right) &\le \mathbb{E}_S\left[w_0(X)\mathbf{1}_{f^\star(X)-\delta\le Y^2\le f^\star(X)}\right] + W\sqrt{\frac{t}{n_S}} \\ &= \mathbb{P}_T\left(f^\star(X) - \delta \le Y^2 \le f^\star(X)\right) + W\sqrt{\frac{t}{n_S}} \\ &\le L\delta + W\sqrt{\frac{t}{n_S}},\end{aligned}$$

where L is upper bound on the density of $Y^2 - f^*(X)$. Call this event Ω_1 that the above bound holds. At this event we have:

$$\begin{aligned}&\frac{1}{n_S/2}\sum_{i\in\mathcal{D}_{S,1}} \hat{w}(X_i)h_\delta\left(Y_i^2 - f^\star(X_i)\right) \\ &= \frac{1}{n_S/2}\sum_{i\in\mathcal{D}_{S,1}} w_0(X_i)h_\delta\left(Y_i^2 - f^\star(X_i)\right) + \frac{1}{n_S/2}\sum_{i\in\mathcal{D}_{S,1}} (\hat{w}(X_i) - w_0(X_i))h_\delta\left(Y_i^2 - f^\star(X_i)\right) \\ &\le L\delta + W\sqrt{\frac{t}{n_S}} + \frac{B+\delta}{\delta}\cdot\frac{2}{n_S}\sum_{i=1}^{n_S/2}|\hat{w}(X_i) - w_0(X_i)|,\end{aligned}$$

where the last inequality follows from the fact that $h_\delta(t) \le (B+\delta)/\delta$ if $t \le B$. Finally, to bound the last summand, we again apply Hoeffding's inequality. As $\|\hat{w}\|_\infty \le W'$, we have with probability greater than or equal to $1 - e^{-t}$:

$$\frac{1}{n_S/2}\sum_{i=1}^{n_S/2}|\hat{w}(X_i) - w_0(X_i)| \le \mathbb{E}_S\left[|\hat{w}(X) - w_0(X)|\right] + (W + W')\sqrt{\frac{t}{n_S}}.$$

If we denote the event Ω_2 where the above inequality holds, then on the event $\Omega_1 \cap \Omega_2$, we have:

$$\begin{aligned}&\frac{1}{n_S/2}\sum_{i} \hat{w}(X_i)h_\delta\left(Y_i^2 - f^\star(X_i)\right) \\ &\le L\delta + W\sqrt{\frac{t}{n_S}} + \frac{B+\delta}{\delta}\cdot\left(\mathbb{E}_S\left[|\hat{w}(X) - w_0(X)|\right] + (W + W')\sqrt{\frac{t}{n_S}}\right) \le \epsilon\,.\end{aligned}$$

Furthermore,

$$\mathbb{P}(\Omega_1 \cap \Omega_2) \geq \mathbb{P}(\Omega_1) + \mathbb{P}(\Omega_2) - 1 \geq 1 - 2e^{-t}.$$

Therefore, we conclude that with probability $\geq 1 - 2e^{-t}$, f^* is a feasible solution.

We now proof Theorem 2.2 on the event $\Omega_1 \cap \Omega_2$, when f^* is a feasible solution. Then we have, $\mathbb{P}_{n,T}(\hat{f}_{\text{init}}(X)) \leq \mathbb{P}_{n,T}(f^*(X))$ on this event, by the optimality of $\hat{f}_{\text{init}}$ in equation (3.5). Then we have:

$$\begin{aligned}
\mathbb{E}_T[\hat{f}_{\text{init}}(X)] &= \mathbb{P}_{n_T}(\hat{f}_{\text{init}}(X)) + (\mathbb{P}_T - \mathbb{P}_{n_T})(\hat{f}_{\text{init}}(X)) \\
&\leq \mathbb{P}_{n_T}(f^*(X)) + (\mathbb{P}_T - \mathbb{P}_{n_T})(\hat{f}_{\text{init}}(X)) \\
&= \mathbb{E}_T[f^*(X)] + (\mathbb{P}_{n_T} - \mathbb{P}_T)(f^*(X) - \hat{f}_{\text{init}}(X)) \\
&\leq \mathbb{E}_T[f^*(X)] + \sup_{f \in \mathcal{F}} |(\mathbb{P}_{n_T} - \mathbb{P}_T)(f^*(X) - f(X))|
\end{aligned}$$

Finally as $f - f^*$ is upper bounded by $F' = B_{\mathcal{F}} + \|f^*\|_\infty$ (as f is uniformly upper bounded by F). Therefore, by Mcdiarmid's inequality, we have with probability $1 - e^{-t}$:

$$\sup_{f \in \mathcal{F}} |(\mathbb{P}_{n_T} - \mathbb{P}_T)(f^*(X) - f(X))| \leq \mathbb{E}_T\left[\sup_{f \in \mathcal{F}} |(\mathbb{P}_{n_T} - \mathbb{P}_T)(f^*(X) - f(X))|\right] + F'\sqrt{\frac{t}{2n_T}}.$$

Call this event Ω_3. Furthermore, by standard symmetrization:

$$\mathbb{E}_T\left[\sup_{f \in \mathcal{F}} |(\mathbb{P}_{n_T} - \mathbb{P}_T)(f^*(X) - f(X))|\right] \leq 2\mathcal{R}_{n_T}(\mathcal{F} - f^*),$$

where $\mathcal{R}_{n_T}(\mathcal{F} - f^*)$ is the Rademacher complexity of $\mathcal{F} - f^*$. Therefore, on $\cap_{i=1}^3 \Omega_i$, we have:

$$\mathbb{E}_T[\hat{f}_{\text{init}}(X)] \leq \mathbb{E}_T[f^*(X)] + 2\mathcal{R}_{n_T}(\mathcal{F} - f^*) + F'\sqrt{\frac{t}{2n_T}},$$

and $\mathbb{P}(\cap_{i=1}^3 \Omega_i) \geq 1 - 3e^{-t}$. This completes the proof.

A.2 Proof of Lemma 3.3

We prove the lemma into two steps; first we show that $\hat{f}_{\text{init}}$ satisfies $\mathbb{P}_T(Y^2 > \hat{f}_{\text{init}}(X)) \leq \tau$ with high probability for some small τ. Next we argue that, on $\mathcal{D}_{S,2}$, we have $(2/n_S) \cdot \sum_{i \in \mathcal{D}_{S,2}} \hat{w}(X_i)\mathbb{1}(Y_i^2 \geq \hat{f}_{\text{init}}(X_i)) \leq \check{\tau}$ with high probability for some small $\check{\tau}$. Then as long as $\check{\tau} \leq \alpha$, we conclude the proof of the lemma.

Step 1: Note that, by feasibility, $\hat{f}_{\text{init}}$ satisfies:

$$\frac{1}{n_S/2} \sum_{i \in \mathcal{D}_{S,1}} \hat{w}(X_i) h_\delta(Y_i^2 - \hat{f}_{\text{init}}(X_i)) \leq \epsilon.$$

This implies:

$$\begin{aligned}
&\mathbb{E}_T\left[h_\delta\left(Y^2 - \hat{f}_{\text{init}}(X)\right)\right] \\
&= \mathbb{E}_S\left[w_0(X) h_\delta\left(Y^2 - \hat{f}_{\text{init}}(X)\right)\right] \\
&= \frac{1}{n_S/2} \sum_{i \in \mathcal{D}_{S,1}} w_0(X_i) h_\delta(Y_S^2 - \hat{f}_{\text{init}}(X_i)) + \left(\mathbb{P}_S - \mathbb{P}_{n_S/2}\right) w_0(X) h_\delta(Y^2 - \hat{f}_{\text{init}}(X)) \\
&= \frac{1}{n_S/2} \sum_{i \in \mathcal{D}_{S,1}} \hat{w}(X_i) h_\delta(Y_i^2 - \hat{f}_{\text{init}}(X_i)) + \frac{1}{n_S/2} \sum_{i \in \mathcal{D}_{S,1}} (w_0(X_i) - \hat{w}(X_i)) h_\delta(Y_i^2 - \hat{f}_{\text{init}}(X_i)) \\
&\qquad + \left(\mathbb{P}_S - \mathbb{P}_{n_S/2}\right) w_0(X) h_\delta(Y^2 - \hat{f}_{\text{init}}(X)) \\
&\leq \epsilon + \frac{B + \delta}{\delta} \|\hat{w} - w_0\|_{L_1(\mathbb{P}_{n_1,S})} + \sup_{f \in \mathcal{F}} \left|\left(\mathbb{P}_S - \mathbb{P}_{n_S/2}\right) w_0(X) h_\delta(Y^2 - f(X))\right|
\end{aligned}$$

https://doi.org/10.52202/079017-2341

Now, as $h_\delta(Y^2 - f(X)) \le (B+\delta)/\delta$ and $w_0 \le W$, we have by Mcdiarmid's inequality, with probability $\ge 1 - e^{-t}$:

$$\begin{aligned}
&\sup_{f\in\mathcal{F}} \left|\left(\mathbb{P}_S - \mathbb{P}_{n_S/2}\right) w_0(X) h_\delta(Y^2 - f(X))\right| \\
&\le \mathbb{E}_S\left[\sup_{f\in\mathcal{F}} \left|\left(\mathbb{P}_S - \mathbb{P}_{n_S/2}\right) w_0(X) h_\delta(Y^2 - f(X))\right|\right] + W\frac{B+\delta}{\delta}\sqrt{\frac{t}{n_S}} \\
&\le 2\mathcal{R}_{n_S/2,\mathcal{F}}(w_0 h_\delta \circ f) + W\frac{B+\delta}{\delta}\sqrt{\frac{t}{n_S}}\,.
\end{aligned}$$

Meanwhile, as in the proof of Theorem 3.2, with probability $\ge 1 - e^{-t}$:

$$\|\hat{w} - w_0\|_{L_1(\mathbb{P}_{n_1,S})} \le \mathbb{E}_S\left[|\hat{w}(X) - w(X)|\right] + (W + W')\sqrt{\frac{t}{n_S}}.$$

Choosing $t = 10\log n_S$ we obtain that with probability $\ge 1 - 2n_S^{-10}$:

$$\begin{aligned}
&\mathbb{E}_T\left(h_\delta\left(Y_T^2 - \hat{f}_{\text{init}}(X_T)\right)\right) \\
&\le \epsilon + \frac{B+\delta}{\delta}\left(\mathbb{E}_S\left[|\hat{w}(X) - w_0(X)|\right] + (W+W')\sqrt{\frac{10\log n_S}{n_S}}\right) \\
&\quad + 2\mathcal{R}_{n_S/2,\mathcal{F}}(w_0 h_\delta \circ f) + W\frac{B+\delta}{\delta}\sqrt{\frac{10\log n_S}{n_S}} \\
&\le \epsilon + \frac{B+\delta}{\delta}\left(\mathbb{E}_S\left[|\hat{w}(X) - w_0(X)|\right] + (2W+W')\sqrt{\frac{10\log n_S}{n_S}}\right) + 2\mathcal{R}_{n_S/2,\mathcal{F}}(w_0 h_\delta \circ f)\,.
\end{aligned}$$

We next bound the Rademacher complexity of $\mathcal{R}_{n_S/2,\mathcal{F}}(w_0 h_\delta \circ f)$. By symmetrization, we have with $\zeta_1, \ldots \zeta_{n_S/2}$ i.i.d. Rademacher$(1/2)$:

$$\begin{aligned}
\mathcal{R}_{n_S/2,\mathcal{F}}(w_0 h_\delta \circ f) &= 2\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n_S/2}\sum_i \zeta_i w_0(X_i) h_\delta(Y_i^2 - f(X_i))\right|\right] \\
&= 2\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n_S/2}\sum_i \zeta_i \phi\left(w_0(X_i), Y_i^2 - f(X_i)\right)\right|\right] \quad [\phi(x,y) = x h_\delta(y)]
\end{aligned}$$

We first show that $\phi : \mathbb{R}^2 \to \mathbb{R}$ is a Lipschitz function on its domain. The first argument of ϕ is $w_0(x)$ which lies within $[-W, W]$. The second argument of ϕ is $Y^2 - f(X)$ (on the source domain), which is bounded by B. Therefore, $h_\delta(Y^2 - f(X))$ is bounded above by $(B+\delta)/\delta$. The derivative of h_δ is 0 for $x \le -\delta$ and δ for $x \ge -\delta$. Hence, we have the following:

$$\|\nabla\phi(x,y)\| = \|(h_\delta(y) \quad x h'_\delta(y))\| \le \sqrt{\frac{(B+\delta)^2}{\delta^2} + \frac{W^2}{\delta^2}} \le \frac{B+W+\delta}{\delta}\,.$$

We next apply vector-valued Ledoux-Talagrand contraction inequality on the function ϕ (equation (1) of [Mau16]), to obtain the following bound on the Rademacher complexity:

$$\begin{aligned}
&2\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n_S/2}\sum_i \zeta_i \phi\left(w_0(X_i), Y_i^2 - f(X_i)\right)\right|\right] \\
&\le 2\sqrt{2}\left(\frac{B+W+\delta}{\delta}\right)\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n_S/2}\sum_i \left(\zeta_{i1} w_0(X_i) + \zeta_{i2}(Y_i^2 - f(X_i))\right)\right|\right] \\
&\le 2\sqrt{2}\left(\frac{B+W+\delta}{\delta}\right)\left[\mathbb{E}_S\left[\left|\frac{1}{n_S/2}\sum_i \zeta_{i1} w_0(X_i)\right|\right] + \mathbb{E}_S\left[\left|\frac{1}{n_S/2}\sum_{i\in\mathcal{D}_{S,1}} \zeta_{i,2} Y_i^2\right|\right]\mathcal{R}_{n_S/2}(\mathcal{F})\right] \\
&\le 2\sqrt{2}\left(\frac{B+W+\delta}{\delta}\right)\left[\frac{\|w_0\|_{L_2(P_{X_S})}}{\sqrt{n_S/2}} + \sqrt{\frac{\mathbb{E}_S[Y^4]}{n_S/2}} + \mathcal{R}_{n_S/2}(\mathcal{F})\right]
\end{aligned}$$

Using this, we obtain the following:

$$
\begin{aligned}
&\mathbb{E}_T\left(h_\delta\left(Y^2-\hat{f}_{\text{init}}(X)\right)\right)\\
&\leq \epsilon+\frac{B+\delta}{\delta}\left(\mathbb{E}_S\left[|\hat{w}(X)-w_0(X)|\right]+(2W+W')\sqrt{\frac{5\log\left(n_S/2\right)}{n_S/2}}\right)\\
&\qquad+4\sqrt{2}\left(\frac{B+W+\delta}{\delta}\right)\left[\frac{\|w_0\|_{L_2(P_{X_S})}+\sqrt{\mathbb{E}_S[Y^4]}}{\sqrt{n_S}}\ +\ \mathcal{R}_{n_S/2}(\mathcal{F})\right]\\
&\leq \epsilon+4\sqrt{2}\left(\frac{B+W+\delta}{\delta}\right)\left[\mathbb{E}\left[|\hat{w}(X_S)-w(X_S)|\right]+(2W+W')\sqrt{\frac{5\log\left(n_S/2\right)}{n_S/2}}\right.\\
&\qquad\left.+\frac{\|w_0\|_{L_2(P_{X_S})}+\sqrt{\mathbb{E}_S[Y^4]}}{\sqrt{n_S/2}}\ +\ \mathcal{R}_{n_S/2}(\mathcal{F})\right]\\
&\leq \epsilon+4\sqrt{2}\left(\frac{B+W+\delta}{\delta}\right)\left[\mathbb{E}\left[|\hat{w}(X_S)-w(X_S)|\right]+(2W+W')\sqrt{\frac{5\log\left(n_S/2\right)}{n_S/2}}+\frac{W+\sqrt{\mathbb{E}_S[Y^4]}}{\sqrt{n_S/2}}\ +\ \mathcal{R}_{n_S/2}(\mathcal{F})\right]
\end{aligned}
$$

Choosing

$$
\epsilon=L\delta+W\sqrt{\frac{5\log\left(n_S/2\right)}{n_S/2}}+\frac{B+\delta}{\delta}\cdot\left(\mathbb{E}_S\left[|\hat{w}(X)-w_0(X)|\right]+(W+W')\sqrt{\frac{5\log\left(n_S/2\right)}{n_S/2}}\right),
$$

we obtain

$$
\begin{aligned}
&\mathbb{E}_T\left(h_\delta\left(Y^2-\hat{f}_{\text{init}}(X)\right)\right)\\
&\lesssim L\delta+\frac{B+W+\delta}{\delta}\cdot\left(\left(\mathbb{E}_S\left[|\hat{w}(X)-w_0(X)|\right]+(W+W')\sqrt{\frac{5\log n_S}{n_S}}+\mathcal{R}_{n_S/2}(\mathcal{F})\right)\right.\\
&\lesssim \sqrt{L(B+W)\left(\left(\mathbb{E}_S\left[|\hat{w}(X)-w_0(X)|\right]+(W+W')\sqrt{\frac{5\log n_S}{n_S}}+\mathcal{R}_{n_S/2}(\mathcal{F})\right)\right.}\\
&\quad+\left(\mathbb{E}_S\left[|\hat{w}(X)-w_0(X)|\right]+(W+W')\sqrt{\frac{5\log n_S}{n_S}}+\mathcal{R}_{n_S/2}(\mathcal{F})\right.\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(by choosing } \delta \text{ to balance the terms)}\\
&\triangleq \tau
\end{aligned}
$$

Call the above event Ω_1. This completes the proof of Step 1.

Step 2: Coming back to $\mathcal{D}_{S,2}$, we have:

$$
\frac{1}{n_S/2}\sum_{i\in\mathcal{D}_{S,2}}\hat{w}(X_{S,i})\mathbb{1}_{Y_i^2>\hat{f}_{\text{init}}(X_i)}\leq\frac{1}{n_S/2}\sum_{i\in\mathcal{D}_{S,2}}|\hat{w}(X_i)-w_0(X_i)|+\frac{1}{n_S/2}\sum_{i\in\mathcal{D}_{S,2}}w_0(X_i)\mathbb{1}_{Y_i^2>\hat{f}_{\text{init}}(X_i)}
$$

Furthermore, by Hoeffding's inequality, we have with probability $\geq 1-e^{-t}$:

$$
\begin{aligned}
\frac{1}{n_S/2}\sum_{i\in\mathcal{D}_2}w_0(X_i)\mathbb{1}_{Y_i^2>\hat{f}_{\text{init}}(X_i)}&\leq\mathbb{E}_S\left[w_0(X)\mathbb{1}_{Y^2>\hat{f}_{\text{init}}(X)}\right]+W\sqrt{\frac{t}{n_S}}\\
&\leq\mathbb{E}_S\left[w_0(X)h_\delta\left(Y^2-\hat{f}_{\text{init}}(X)\right)\right]+W\sqrt{\frac{t}{n_S}}\\
&=\mathbb{E}_T\left(h_\delta\left(Y^2-\hat{f}_{\text{init}}(X)\right)\right)+W\sqrt{\frac{t}{n_S}}
\end{aligned}
$$

Meanwhile, with probability $\geq 1-e^{-t}$:

$$
\frac{1}{n_S/2}\sum_{i\in\mathcal{D}_{S,2}}|\hat{w}(X_i)-w_0(X_i)|\leq\mathbb{E}_S\left[|\hat{w}(X)-w_0(X)|\right]+(W+W')\sqrt{\frac{t}{n_S}}.
$$

https://doi.org/10.52202/079017-2341

Therefore, with $t = 10 \log n_S$, we have with probability $\geq 1 - 2n_S^{-10}$:

$$\frac{1}{n_S/2} \sum_{i \in \mathcal{D}_{S,2}} \hat{w}(X_i) \mathbb{1}_{Y_i^2 > \hat{f}_{\text{init}}(X_i)} \leq \mathbb{E}_S \left[|\hat{w}(X) - w_0(X)| \right] + (W + W') \sqrt{\frac{10 \log n_S}{n_S}} + \mathbb{E}_T \left(h_\delta \left(Y^2 - \hat{f}_{\text{init}}(X) \right) \right) + W \sqrt{\frac{10 \log n_S}{n_S}} .$$

Call this event Ω_2. Therefore, on $\Omega_1 \cap \Omega_2$ we have:

$$\frac{1}{n_S/2} \sum_{i \in \mathcal{D}_{S,2}} \hat{w}(X_i) \mathbb{1}_{Y_i^2 > \hat{f}_{\text{init}}(X_i)} \leq \mathbb{E}_S \left[|\hat{w}(X) - w_0(X)| \right] + (2W + W') \sqrt{\frac{10 \log n_S}{n_S}} + \tau \triangleq \tilde{\tau} .$$

This completes the proof of Step 2. For any fixed $\alpha > 0$, we have $\tilde{\tau} \leq \alpha$ as long as n_S is large enough and $\mathbb{E}_S \left[|\hat{w}(X) - w_0(X)| \right]$ is small enough, and as a consequence $\hat{\lambda}(\alpha) \leq 1$. This completes the proof.

A.3 Proof of Theorem 3.4

Recall that we construct the prediction intervals using data splitting; from the first part of the data (namely $\mathcal{D}_1$), we estimate $\hat{f}_{\text{init}}$ and use the second part of the data (namely $\mathcal{D}_2$) to estimate $\hat{\lambda}(\alpha)$. Conditional on $\mathcal{D}_1$, define a function class $\mathcal{G} \equiv \mathcal{G}(\hat{f})$ as:

$$\mathcal{G} = \left\{ g_\lambda(x, y) = w_0(x) \mathbb{1}_{y^2 - \lambda \hat{f}_{\text{init}}(x) \geq 0} : \lambda \geq 0 \right\} .$$

As $\mathcal{G}$ only depends on a scalar parameter λ (as w_0 and $\hat{f}_{\text{init}}$ are fixed conditionally on $\mathcal{D}_{S,1}, \mathcal{D}_T$), it is a VC class of function with VC-dim ≤ 2.

$$\begin{aligned}
&\mathbb{P}_T \left(Y^2 \geq \hat{\lambda}(\alpha) \hat{f}_{\text{init}}(X) \right) \\
&= \mathbb{E}_S \left[w_0(X) \mathbb{1}_{Y^2 - \hat{\lambda}(\alpha) \hat{f}_{\text{init}}(X) \geq 0} \right] \\
&= \frac{1}{n_S/2} \sum_{i \in \mathcal{D}_{S,2}} w_0(X_i) \mathbb{1}_{Y_i^2 - \hat{\lambda}(\alpha) \hat{f}_{\text{init}}(X_i)} + (\mathbb{P}_S - \mathbb{P}_{n_S/2}) w_0(X) \mathbb{1}_{Y^2 \geq \hat{\lambda}(\alpha) \hat{f}_{\text{init}}(X)} \\
&= \frac{1}{n_S/2} \sum_{i \in \mathcal{D}_{S,2}} \hat{w}(X_i) \mathbb{1}_{Y_i^2 - \hat{\lambda}(\alpha) \hat{f}_{\text{init}}(X_i) \geq 0} + \frac{1}{n_S/2} \sum_{i \in \mathcal{D}_{S,2}} (w_0(X_i) - \hat{w}(X_i)) \mathbb{1}_{Y_i^2 - \hat{\lambda}(\alpha) \hat{f}_{\text{init}}(X_i) \geq 0} \\
&\qquad + (\mathbb{P}_S - \mathbb{P}_{n_S/2}) w_0(X) \mathbb{1}_{Y^2 - \hat{\lambda}(\alpha) \hat{f}_{\text{init}}(X) \geq 0} \qquad \text{(A.1)}
\end{aligned}$$

Now, by the definition of $\hat{\lambda}(\alpha)$ (see Step 2), we have:

$$\alpha - \frac{1}{n_S/2} \leq \frac{1}{n_S/2} \sum_{i \in \mathcal{D}_{S,2}} \hat{w}(X_i) \mathbb{1}_{Y_i^2 - \hat{\lambda}(\alpha) \hat{f}_{\text{init}}(X_i) \geq 0} \leq \alpha .$$

We use a similar technique to control the second summand as in the proof of Theorem 3.2. By using the fact that the indicator function is less than one, we have:

$$\left| \frac{1}{n_S/2} \sum_{i \in \mathcal{D}_{S,2}} (w_0(X_i) - \hat{w}(X_i)) \mathbb{1}_{Y_i^2 - \hat{\lambda}(\alpha) \hat{f}_{\text{init}}(X_i) \geq 0} \right| \leq \frac{1}{n_S/2} \sum_{i \in \mathcal{D}_{S,2}} |\hat{w}(X_i) - w_0(X_i)| .$$

Applying Hoeffding's inequality (with the fact that $\|\hat{w}\|_\infty \leq W'$ and $\|w_0\|_\infty \leq W$), we have with probability greater than or equal to $1 - e^{-t}$:

$$\frac{1}{n_S/2} \sum_{i \in \mathcal{D}_{S,2}} |\hat{w}(X_i) - w_0(X_i)| \leq \mathbb{E}_S \left[|\hat{w}(X) - w(X)| \right] + (W + W') \sqrt{\frac{t}{n_S}} .$$

To control the third summand of (A.1), note that, conditional on $\mathcal{D}_{S,1}$ and $\mathcal{D}_T$ (i.e., assuming $\hat{f}_{\text{init}}$ fixed), and using the fact that $\|g\|_\infty \leq \|w_0\|_\infty \leq W$ for all $g \in \mathcal{G}$, we have by Mcdiarmid's

inequality with probability greater than or equal to $1 - e^{-t}$:

$$\sup_{g\in\mathcal{G}} \left|(\mathbb{P}_S - \mathbb{P}_{n_S/2})g(X,Y)\right| \le \mathbb{E}_S\left[\sup_{g\in\mathcal{G}} \left|(\mathbb{P}_S - \mathbb{P}_{n_S/2})g(X,Y)\right| \mid \mathcal{D}_{S,1}, \mathcal{D}_T\right] + W\sqrt{\frac{t}{n_S}}$$
$$\le 2\mathcal{R}_{n_S/2}\left(\mathcal{G} \mid \mathcal{D}_{S,1}, \mathcal{D}_T\right) + W\sqrt{\frac{t}{n_S}}.$$

Now conditional on $\mathcal{D}_{S,1}, \mathcal{D}_T$, $\mathcal{G}$ is a VC class of function with VC dimension ≤ 2. Therefore,

$$\mathcal{R}_{n_S/2}\left(\mathcal{G} \mid \mathcal{D}_{S,1}, \mathcal{D}_T\right) \le \sqrt{\frac{C}{n_S}}$$

for some constant $C > 0$. Thus, we have

$$\sup_{g\in\mathcal{G}} \left|(\mathbb{P}_S - \mathbb{P}_{n_S/2})g(X,Y)\right| \le \sqrt{\frac{C}{n_S}} + W\sqrt{\frac{t}{n_S}}.$$

Combining the bounds, we have, with probability $\ge 1 - 2e^{-t}$:

$$\left|\mathbb{P}_T\left(Y^2 > \hat{\lambda}(\alpha)\hat{f}_{\text{init}}(X)\right) - \alpha\right|$$
$$\le \frac{1}{n_S/2} + \mathbb{E}_S\left[|\hat{w}(X) - w_0(X)|\right] + (2W + W')\sqrt{\frac{t}{n_S}} + \sqrt{\frac{C}{n_S}}.$$

This completes the proof.

A.4 Proof of Theorem 4.3

We start with the following decomposition:

$$\begin{aligned}\mathbb{E}_T[\hat{f}_{\text{init}} \circ \hat{T}_0(X)] &= \mathbb{E}_T[\hat{f}_{\text{init}} \circ T_0(X)] + \mathbb{E}_T[\hat{f}_{\text{init}} \circ \hat{T}_0(X) - \hat{f}_{\text{init}} \circ T_0(X)] \\ &= \mathbb{E}_S[\hat{f}_{\text{init}}(X)] + \mathbb{E}_T[\hat{f}_{\text{init}} \circ \hat{T}_0(X) - \hat{f}_{\text{init}} \circ T_0(X)] \\ &\le \mathbb{E}_S[\hat{f}_{\text{init}}(X)] + L_{\mathcal{F}}\mathbb{E}_T[|\hat{T}_0(X) - T_0(X)|]\end{aligned}$$

where the second equation follows from the fact that when $X \sim P_T$, then $T_0(X) \sim P_S$, and the last line follows from the fact $f \in \mathcal{F}$ is $L_{\mathcal{F}}$ Lipschitz. A similar argument as in the proof of Theorem 3.5 [FGM23] yields:

$$\mathbb{E}_S[\hat{f}_{\text{init}}(X)] \le \Delta + 4\mathcal{R}_{n_S}(\mathcal{F}) + 4B_{\mathcal{F}}\sqrt{\frac{t}{2n_S}}.$$

with probability $\ge 1 - e^{-t}$. We then finish the proofs.

A.5 Proof of Lemma 4.4

By the definition of $\hat{\lambda}(\alpha)$, we have

$$\left\{\hat{\lambda}(\alpha) \ge 1\right\} \implies \left\{\frac{1}{n_S/2}\sum_{i\in\mathcal{D}_{S,2}} \mathbb{1}\left(Y_i^2 \ge \hat{f}_{\text{init}}(X_i) + \delta\right) > \alpha\right\}.$$

Now by an application of Chernoff bound for binomial distribution, we have:

$$\mathbb{P}\left(\frac{1}{n_S/2}\sum_{i\in\mathcal{D}_{S,2}} \mathbb{1}\left(Y_i^2 \ge \hat{f}_{\text{init}}(X_i) + \delta\right) > \alpha \mid \mathcal{D}_{S,1}, \mathcal{D}_T\right) \le e^{-\frac{(\alpha - p_{n_S})^2 n_S}{6p_{n_S}}}.$$

Hence, we have the following:

$$\mathbb{P}(\hat{\lambda}(\alpha) > 1 \mid \mathcal{D}_{S,1}, \mathcal{D}_T) \le e^{-\frac{(\alpha - p_{n_S})^2 n_S}{6p_{n_S}}}.$$

https://doi.org/10.52202/079017-2341

We next establish the high probability bound on p_{n_S}. We define a function $\ell_\delta(x)$ which is 1 when $x \le -\delta$, 0 when $x \ge 0$ and $-x/\delta$ when $-\delta \le x \le 0$.

$$\begin{aligned} p_{n_S} = \mathbb{E}_S\left[\mathbb{1}_{Y^2 \ge \hat{f}_{\text{init}}(X)+\delta}\right] &\le \mathbb{E}_S\left[\ell_\delta(\hat{f}_{\text{init}}(X) - Y^2)\right] \\ &= \frac{1}{n_S/2}\sum_{i \in \mathcal{D}_{S,1}} \ell_\delta(\hat{f}_{\text{init}}(X_i) - Y_i^2) + \left(\mathbb{P}_{n_S/2} - \mathbb{P}_S\right)\ell_\delta(\hat{f}_{\text{init}}(X) - Y^2) \\ &\le \sup_{f \in \mathcal{F}} \left(\mathbb{P}_{n_S/2} - \mathbb{P}_S\right)\ell_\delta(f(X) - Y^2) \\ &\le \frac{4}{\delta}\left(\sqrt{\frac{\mathbb{E}_S[Y^4]}{n_S}} + \mathcal{R}_{n_S/2}(\mathcal{F})\right) + \sqrt{\frac{t}{n_S}}\,. \end{aligned}$$

where the first inequality used $\ell_\delta(x) \ge \mathbb{1}(x \le -\delta)$, second inequality uses the fact that sample average of ℓ_δ over $\mathcal{D}_{S,1}$ is 0 by the definition of $\hat{f}_{\text{init}}$, third inequality uses Ledoux-Talagrand contraction inequality observing that ℓ_δ is $1/\delta$-Lipschitz. This completes the proof.

A.6 Proof of Theorem 4.5

$$\begin{aligned} &\mathbb{P}_T\left(Y^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}} \circ \hat{T}_0(X) + \delta)\right) \\ &= \mathbb{P}_T\left(Y^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}} \circ T_0(X) + \delta)\right) \\ &\qquad + \left|\mathbb{P}_T\left(Y^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}} \circ \hat{T}_0(X) + \delta)\right) - \mathbb{P}_T\left(Y^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}} \circ T_0(X) + \delta)\right)\right| \\ &\triangleq T_1 + T_2\,. \end{aligned} \tag{A.2}$$

We start with analyzing the first term:

$$\begin{aligned} T_1 &= \mathbb{P}_T\left(Y^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}} \circ T_0(X) + \delta)\right) \\ &= \int_{\mathcal{X}_T}\int_{\mathcal{Y}} \mathbb{1}_{y^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(T_0(x))+\delta)}\, f_T(y \mid X_T = x)\, p_T(x)\, dydx \\ &= \int_{\mathcal{X}_T}\int_{\mathcal{Y}} \mathbb{1}_{y^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(T_0(x))+\delta)}\, f_S(y \mid X_S = T_0(x))\, p_T(x)\, dydx \\ &= \int_{\mathcal{X}_S}\int_{\mathcal{Y}} \mathbb{1}_{y^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(z)+\delta)}\, f_S(y \mid X_S = z)\, p_T(T_0^{-1}(z))|\nabla T_0^{-1}(z)|\, dydx \\ &= \int_{\mathcal{X}_S}\int_{\mathcal{Y}} \mathbb{1}_{y^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(z)+\delta)}\, f_S(y \mid X_S = z)\, p_S(z)\, dydx \\ &= \mathbb{P}_S(Y^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(X) + \delta))\,. \end{aligned}$$

Therefore, we need a high probability upper bound on $\mathbb{P}_S(Y^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(X) + \delta) \mid \mathcal{D}_S \cup \mathcal{D}_T)$. Towards that end, we start with the following expansion:

$$\begin{aligned} &\mathbb{P}_S\left(Y^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(X) + \delta) \mid \mathcal{D}_S \cup \mathcal{D}_T\right) \\ &= \frac{1}{n_S/2}\sum_{i \in \mathcal{D}_{S,2}} \mathbb{1}_{Y_i^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(X_i)+\delta)} + \left(\mathbb{P}_{n_S/2} - \mathbb{P}_S\right)\mathbb{1}_{Y^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(X)+\delta)} \end{aligned} \tag{A.3}$$

Now, note that, by the definition of $\hat{\lambda}(\alpha)$, we have:

$$\alpha - \frac{1}{n_S/2} \le \frac{1}{n_S/2}\sum_{i \in \mathcal{D}_{S,2}} \mathbb{1}_{Y_i^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(X_i)+\delta)} \le \alpha\,.$$

To bound the second term in (A.3), we use:

$$\left|\left(\mathbb{P}_{n_S/2} - \mathbb{P}_S\right)\mathbb{1}_{Y^2 \ge \hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(X)+\delta)}\right| \le \sup_{\lambda \ge 0}\left|\left(\mathbb{P}_{n_S/2} - \mathbb{P}_S\right)\mathbb{1}_{Y^2 \ge \lambda(\hat{f}_{\text{init}}(X)+\delta)}\right| := \mathbf{Z}_n\,.$$

To bound the supremum we use standard techniques from the empirical process theory. Define a collection of functions $\mathcal{G} = \left\{ \mathbb{1}_{Y^2 \geq \lambda(\hat{f}_{\text{init}}(X)+\delta)} : \lambda \geq 0 \right\}$. Note that, here we condition on $\mathcal{D}_{S,1}$, so we treat $\hat{f}_{\text{init}}$ as a constant function. For notational simplicity, suppose

$$\Psi_n = \mathbb{E}_S \left[\sup_{\lambda \geq 0} \left| \left(\mathbb{P}_{n_S/2} - \mathbb{P}_S \right) \mathbb{1}_{Y^2 \geq \lambda(\hat{f}_{\text{init}}(X)+\delta)} \right| \mid \mathcal{D}_{S,1} \right] = \mathbb{E}_S \left[\sup_{g \in \mathcal{G}} \left| \left(\mathbb{P}_{n_S/2} - \mathbb{P}_S \right) g(X,Y) \right| \mid \mathcal{D}_{S,1} \right] .$$

As the functions in $\mathcal{G}$ are uniformly bounded by 1 (and consequently, $\mathbb{E}[g^2(X,Y)] \leq 1$), we have by Talagrand's concentration inequality of the suprema of the empirical process:

$$\mathbb{P} \left(\mathbf{Z}_n \geq \Psi_n + \sqrt{2t \frac{1 + 4\Psi_n}{n_S}} + \frac{4t}{3n_S} \mid \mathcal{D}_{S,1} \right) \leq e^{-t} . \tag{A.4}$$

Therefore, we need an upper bound on Ψ_n to obtain a high probability upper bound on $\mathbf{Z}_n$. Towards that end, observe that $\mathcal{G}$ is a VC class with VC-dim less than or equal to 2 (as it is an indicator function of a collection of functions with one parameter). Hence, we have, by symmetrization and Dudley's metric entropy bound:

$$\Psi_n \leq 2\mathbb{E}_S \left[\sup_{g \in \mathcal{G}} \left| \frac{1}{n_S/2} \sum_{i \in \mathcal{D}_{S,2}} \epsilon_i g(X_i, Y_i) \right| \mid \mathcal{D}_{S,1} \right] \leq \frac{C}{\sqrt{n_S}} .$$

Therefore, going back to (A.4), we have with probability $\geq 1 - e^{-t}$

$$\mathbf{Z}_n \leq \frac{C}{\sqrt{n_S}} + \sqrt{\frac{C_1}{n_S} + \frac{C_2}{n_S^{3/2}}} \sqrt{t} + \frac{4t}{3n_S} .$$

Hence, we have:

$$\left| \mathbb{P}_S \left(Y^2 \geq \hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(X) + \delta) \mid \mathcal{D}_S \cup \mathcal{D}_T \right) - \alpha \right| \lesssim \sqrt{\frac{t}{n_S}}$$

with probability $\geq 1 - e^{-t}$. This completes the proof of T_1. To obtain a bound on T_2, note that:

$$\begin{aligned}
&T_2 \\
&= \left| \mathbb{P}_T \left(Y^2 \geq \hat{\lambda}(\alpha)(\hat{f}_{\text{init}} \circ \hat{T}_0(X) + \delta) \right) - \mathbb{P}_T \left(Y^2 \geq \hat{\lambda}(\alpha)(\hat{f}_{\text{init}} \circ T_0(X) + \delta) \right) \right| \\
&= \left| \int_{\mathcal{X}_T} \Big(\mathbb{P}_T(Y^2 \leq \hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(\hat{T}_0(x)) + \delta) \mid X_T = x) \right. \\
&\qquad \left. - \mathbb{P}_T(Y^2 \leq \hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(T_0(x)) + \delta) \mid X_T = x) \Big) \; p_T(x) \, dx \right| \\
&= \left| \int_{\mathcal{X}_T} \left(F_{Y_T^2 | X_T = x}(\hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(\hat{T}_0(x)) + \delta)) - F_{Y_T^2 | X_T = x}(\hat{\lambda}(\alpha)(\hat{f}_{\text{init}}(T_0(x)) + \delta) \right) \; p_T(x) \, dx \right| \\
&\leq G \int_{\mathcal{X}_T} \hat{\lambda}(\alpha) \left| \hat{f}_{\text{init}}(T_0(x)) - \hat{f}_{\text{init}}(\hat{T}_0(x)) \right| \; p_T(x) \, dx \\
&\leq G L_{\mathcal{F}} \, \mathbb{E}_T[|T_0(X) - \hat{T}_0(X)|] .
\end{aligned}$$

Here, the penultimate inequality uses the fact that the conditional distribution of Y_T^2 given X_T is Lipschitz (as the density of Y_T^2 given X_T is bounded), and the last inequality uses the fact that $\hat{f}_{\text{init}}$ is Lipschitz as we have assumed all functions in $\mathcal{F}$ are Lipschitz.

B Details of the experiment

B.1 Density ratio estimation via probabilistic classification

Suppose we observe $\{X_1, \ldots, X_{n_1}\}$ from a distribution P (with density p) and $\{X_{n_1+1}, \ldots, X_{n_1+n_2}\}$ from another distribution Q (with density q). We are interested in estimating $w_0(x) = q(x)/p(x)$, where we assume Q is absolutely continuous with respect to P

https://doi.org/10.52202/079017-2341

(otherwise, the density ratio can be unbounded with positive probability). Define, $n_1 + n_2$ mane binary random variables $\{C_1, \ldots, C_{n_1+n_2}\}$ such that $C_i = 0$ for $1 \leq i \leq n_1$ and $C_i = 1$ for $n_1 + 1 \leq i \leq n_1 + n_2$. Consider the augmented dataset $\mathcal{D} = \{(X_i, C_i)\}_{1 \leq i \leq n_1+n_2}$. We can think that this dataset is generated from a mixture distribution $\rho p(X) + (1-\rho)q(x)$ where $\rho = \mathbb{P}(C = 1)$. For this mixture distribution, the posterior distribution of C given X is:

$$\begin{aligned}\mathbb{P}(C = 1|X = x) &= \frac{P(X = x \mid C = 1)P(C = 1)}{P(X = x \mid C = 1)P(C = 1) + P(X = x \mid C = 0)P(C = 0)} \\ &= \frac{\rho q(x)}{\rho q(x) + (1 - \rho)p(x)} \\ &= \frac{(\rho/(1 - \rho))w_0(x)}{(\rho/(1 - \rho))w_0(x) + 1}\end{aligned}$$

This implies:

$$w_0(x) = \frac{1 - \rho}{\rho} \frac{\mathbb{P}(C = 1 \mid X = x)}{1 - \mathbb{P}(C = 1 \mid X = x)}.$$

Now, from the data, we can estimate $\hat{\rho} = n_2/(n_1 + n_2)$ and $\mathbb{P}(C = 1 \mid X = x)$ by any classification technique (e.g., using logistic regression, boosting, random forest, deep neural networks etc). Let $\hat{g}(x)$ be one such classifier. Then we can estimate $w_0(x)$ by $(n_1/n_2)(\hat{g}(x)/(1 - \hat{g}(x)))$.

B.2 General weighted conformal prediction

The weighted conformal prediction method, as presented in [TFBCR19], consists of two main steps:

1. Split the source data into parts; estimate the conditional mean function $\mathbb{E}[Y \mid X = x]$, say $\hat{\mu}(x)$ using the first part of the source data.
2. Use the second part of the source data and the target data to construct weight $w(X_i)$ and the score function $S(x, y) = |y - \hat{\mu}(x)|$ to construct the confidence interval.

In Section 5, we have implemented a generalized version of it, where we modify the score function as follows:

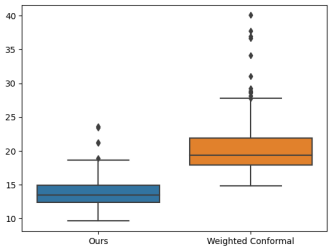
1. We estimate the conditional standard deviation function $\sqrt{\mathsf{var}(Y \mid X = x)}$ along with the conditional mean function from the first part of the data. Call it $\hat{\sigma}(x)$.
2. We use the modified score function $s(x, y) = |y - \hat{\mu}(x)|/\hat{\sigma}(x)$.

The rest of the method is the same as [TFBCR19]. This additional estimated conditional variance function allows more expressivity and flexibility to the conformal prediction band, as observed in Section 5.2 of [LGR+18], as this captures the local heterogeneity of the conditional distribution of Y given X.

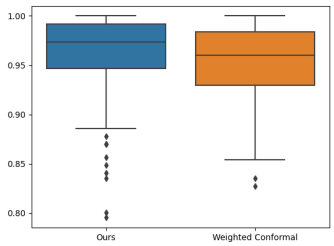
B.3 Boxplots to compare coverage and bandwidth

In this subsection, we present two boxplots to compare the variation in coverage and average width of the prediction bands between our method and the generalized weighted conformal prediction (as described in the previous subsection).

(a) Average Bandwidth



(b) Coverage



Real Estate Data			
Outcome	Our Method	WVAC	WQC
Coverage (Median)	0.98	0.962	0.971
Coverage (IQR)	0.058	0.048	0.048
Bandwidth (Median)	36.889	46.392	46.858
Bandwidth (IQR)	15.001	19.027	14.189
Bandwidth (Median for Coverage $> 95\%$)	40.774	50.703	51.031

Table 2: Experimental results for the real estate data

The boxplots immediately show that our methods yield similar coverage (even with lesser variability) with significantly lower average width than the generalized weighted conformal prediction method.

C Additional Experiments

In this section, we present the details of the additional experiments based on four other datasets, as mentioned in Section 5. We compare our methodology against two other widely used conformal methods, namely Weighted Variance-adjusted Conformal (WVAC) [TFBCR19], and Weighted Quantile-adjusted Conformal (WQC)—which is a variant of WVAC, where the score function is changed to the conditional quantile [RPC19].

C.1 Experiment 1: Real estate data

The Real Estate Valuation dataset available at the UCI Machine Learning Repository contains data used to estimate the value of real estate properties. Collected from Taipei, Taiwan, this dataset is useful for predictive modeling in the real estate sector. It consists of 414 instances, with each entry representing a different real estate transaction. This dataset was originally introduced in [TX12a]. In this dataset, the goal is to predict the house price per unit area based on the 6 other features, namely i) transaction date, ii) house age, iii) distance to the nearest MRT station, iv) number of convenience stores, v) latitude and vi) longitude. The construction of shifted data (with $\beta = (-1, 0, -1, 0, 1, 1)$) and implementation procedure are the same as in Section 5. Table 2 and Figure 3 presented the results over 200 Monte Carlo iterations. It is evident from the table that our method produced a small average width in comparison to the other methods while maintaining the coverage guarantee.

C.2 Experiment 2: Energy efficiency data

The Energy Efficiency dataset available at the UCI Machine Learning Repository is designed to help predict the heating load and cooling load requirements of buildings. The dataset was originally introduced in [YH18]. The dataset includes various building parameters such as: i) wall area, ii) surface area, iii) roof area, iv) orientation, etc. The goal is to predict the heating load based on 8 other covariates. The construction of shifted data (with $\beta = (-1, 0, 1, 0, -1, 0, 0, -1)$) and implementation procedure are the same as in Section 5. Our results over 200 Monte Carlo iterations are presented in Table 3 and Figure 4. The table shows that our method produced a smaller bandwidth than WVAC. While WQC has a smaller median bandwidth, it sacrifices coverage. The last row indicates that for experiments with coverage $\geq 95\%$, WQC's median average width is significantly larger than ours. Thus, whenever WQC provides adequate coverage, its bandwidth is much larger than ours.

C.3 Experiment 3: Appliances Energy Prediction Dataset

Appliances Energy Prediction Dataset is freely available from the UCI repository. This dataset is a time series data with 28 covariates and one response variable. We used data from 2016-01-11 to 2016-02-15 (5000 samples) as our training set and data from 2016-05-13 to 2016-05-27 (2000

https://doi.org/10.52202/079017-2341

Energy efficiency data			
Outcome	Our Method	WVAC	WQC
Coverage (Median)	0.995	0.969	0.973
Coverage (IQR)	0.047	0.036	0.05
Bandwidth (Median)	4.332	5.045	2.842
Bandwidth (IQR)	1.358	3.269	2.551
Bandwidth (Median for Coverage $> 95\%$)	4.373	5.681	4.94

Table 3: Experimental results for the energy efficiency data

samples) as our testing set. Since the source and target data are from different time periods, this experiment involves a non-synthetic real-world time shift. The results based on our Algorithm 2 are presented below:

Appliances Energy Prediction Dataset			
Outcome	Our Method	WVAC	WQC
Coverage	0.95	1.00	1.00
Bandwidth	461.69	6809.87	2032.12

Table 4: Experimental results for the Appliances Energy Prediction Dataset

C.4 Experiment 4: ETDataset

We applied our method to the ETDataset (ETT-small) from [ZZP+21], which contains hourly-level data from two electricity transformers at two different stations, including load and oil temperature measurements. Each data point consists of 8 features, including the date of the point, the predictive value "oil temperature", and 6 different types of external power load features. For our experiment, we used the data from one transformer during the period from July 1, 2016, to November 2, 2016, as our source data, and data from the same time period from the other transformer as our target data. As our source data and the target data are from different locations, we have a geographical covariate shift; see [ZZP+21] for more details. Our results are as follows:

ETDataset			
Outcome	Our Method	WVAC	WQC
Coverage	0.976	0.982	0.842
Bandwidth	41.525	57.9	54.981

Table 5: Experimental results for the ETDataset

C.5 Experiment 5: Airfoil data

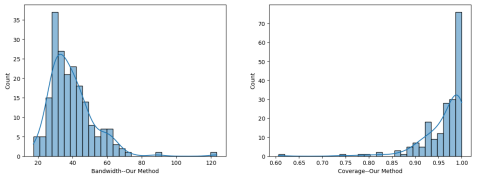
In Section 5, we have implemented our Algorithm 1 on the Airfoil data [TFBCR19] to showcase the efficacy of our method. Here, additionally, we implement Algorithm 2 on the same dataset to evaluate the performance of our second method based on optimal transport-based domain alignment. Here, the data shifting procedure (to create data from the target domain with a shifted distribution) is different from the rest of the experiments; we use the linear transformation $x \mapsto Ax + b$ with $A = \mathsf{diag}(1.5, 1.2, 1.6, 2, 1.8)$ and $b = (1, 0, 0, 1, 0)$ to generate covariates on the target domain. As before, we split the data into two parts; we keep 75% of the data as it is, and shift the rest 25% of the data. The results are given in Table 6 and Figure 5. The second column of Table 6, namely *Our Method (Without OT)* is the aggregation method proposed in this paper, without domain alignment

Airfoil data				
Outcome	Our Method	Our Method (without OT)	WVAC	WQC
Coverage (Median)	0.928	0.749	0.984	0.952
Coverage (IQR)	0.035	0.22	0.024	0.077
Bandwidth (Median)	15.075	18.512	36.298	32.143
Bandwidth (IQR)	1.638	3.089	10.619	8.364
Bandwidth (Median for Coverage $> 95\%$)	16.429	25.268	37.783	36.433

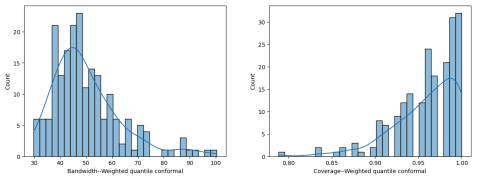
Table 6: Experimental results for the airfoil data using optimal transport

through optimal transport, i.e., we use source data to construct the prediction interval and use the same prediction interval on the target domain. As evident from Table 6, our method outperforms all other methods in terms of the average width of the prediction interval while maintaining a good coverage guarantee.

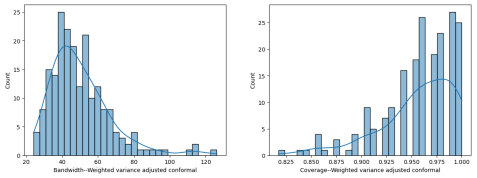
https://doi.org/10.52202/079017-2341



(a) Our-Width (b) Our-Coverage

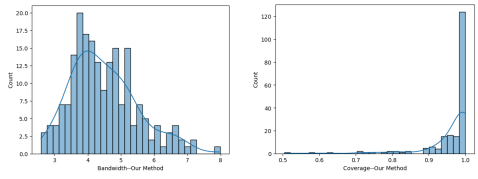


(c) WQC-Width (d) WQC-Coverage

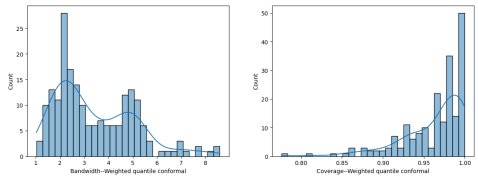


(e) WVAC-Width (f) WVAC-Coverage

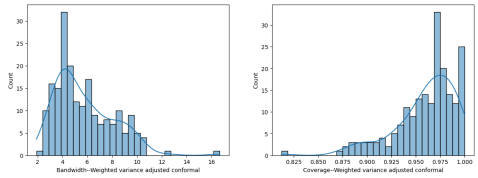
Figure 3: Experiments on real estate data



(a) Our-Width (b) Our-Coverage

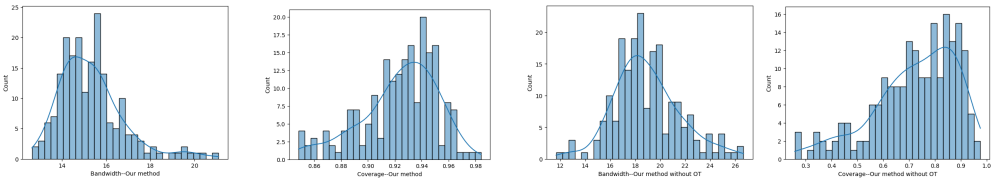


(c) WQC-Width (d) WQC-Coverage

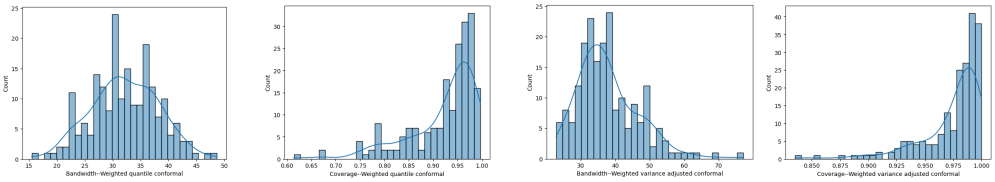


(e) WVAC-Width (f) WVAC-Coverage

Figure 4: Experiments on Energy efficiency data



(a) Our-Width (b) Our-Coverage (c) Our(No OT)-Width (d) Our(No OT)-Coverage



(e) WQC-Width (f) WQC-Coverage (g) WVAC-Width (h) WVAC-Coverage

Figure 5: Experiments on Airfoil data using optimal transport

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See Section 3, 4 and 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

   Justification: We have proposed a method for aggregating prediction intervals along with theoretical guarantees and experiments in this paper. Our theory is valid under certain assumptions clearly mentioned in the main paper. One limitation of the theory is that we do not know the performance of our method if the assumptions are violated, as is true for any theory of methods. Furthermore, one needs to verify the methodology on various real datasets to under the applicability of our method, which is outside the scope of this paper. As these are quite standard limitations, we refrain from including them in the main draft. However, if the referees feel that this should be included in the main draft, we would be happy to oblige.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

https://doi.org/10.52202/079017-2341

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: See Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: See Section 5.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

https://doi.org/10.52202/079017-2341

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Both methods in the experiment took approximately 10 minutes to run on a MacBook Pro laptop (with M2 Max CPU, 10 Cores, 32 GB RAM, and no GPU).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

Justification: This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: See Section 5.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

https://doi.org/10.52202/079017-2341

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2341