The Annals of Statistics
2024, Vol. 52, No. 4, 1796–1824
https://doi.org/10.1214/24-AOS2422

EFFICIENT AND MULTIPLY ROBUST RISK ESTIMATION UNDER GENERAL FORMS OF DATASET SHIFT

BY HONGXIANG QIU[1,a], ERIC TCHETGEN TCHETGEN[2,b] AND EDGAR DOBRIBAN[2,c]

[1]*Department of Epidemiology and Biostatistics, Michigan State University,* [a]*qiuhongx@msu.edu*
[2]*Department of Statistics and Data Science, the Wharton School, University of Pennsylvania,* [b]*ett@wharton.upenn.edu,* [c]*dobriban@wharton.upenn.edu*

Statistical machine learning methods often face the challenge of limited data available from the population of interest. One remedy is to leverage data from auxiliary source populations, which share some conditional distributions or are linked in other ways with the target domain. Techniques leveraging such *dataset shift* conditions are known as *domain adaptation* or *transfer learning*. Despite extensive literature on dataset shift, limited works address how to efficiently use the auxiliary populations to improve the accuracy of risk evaluation for a given machine learning task in the target population.

In this paper, we study the general problem of efficiently estimating target population risk under various dataset shift conditions, leveraging semiparametric efficiency theory. We consider a general class of dataset shift conditions, which includes three popular conditions—covariate, label and concept shift—as special cases. We allow for partially nonoverlapping support between the source and target populations. We develop efficient and multiply robust estimators along with a straightforward specification test of these dataset shift conditions. We also derive efficiency bounds for two other dataset shift conditions, posterior drift and location-scale shift. Simulation studies support the efficiency gains due to leveraging plausible dataset shift conditions.

1. Introduction.

1.1. *Background.* A common challenge in statistical machine learning approaches to prediction is that limited data is available from the population of interest, despite potentially large amounts of data from similar populations. For instance, it may be of interest to predict HIV treatment response in a community based on only a few observations. A large dataset from another community in a previous HIV study may help improve the training of such a prediction model. Another example is building a classification or diagnosis model based on medical images for lung diseases (Christodoulidis et al. (2017)). A key task therein is to classify the texture in the image, but the size of a labeled medical image sample is often limited due to the high cost of data acquisition and labeling. It may be helpful to leverage large existing public image datasets as supplemental data to train the classifier.

In these examples and others, it is desirable to use data from similar source populations to supplement target population data, under plausible *dataset shift* conditions relating the source and target populations (see, e.g., Storkey (2013), Shimodaira (2000), Sugiyama and Kawanabe (2012)). Such methods are known as *domain adaptation* or *transfer learning* (see, e.g., Kouw and Loog (2018), Pan and Yang (2010)).

A great deal of work has been devoted to leveraging or addressing various dataset shift conditions. Polo et al. (2022) proposed testing for various forms of dataset shift. Among dataset

Received March 2024; revised June 2024.
MSC2020 subject classifications. 68Q32, 62G20.
Key words and phrases. Dataset shift, domain adaptation, efficiency, multiple robustness, semiparametric model, transfer learning.

shift types, popular conditions include *covariate shift*, where only the covariate distribution changes, as well as *label shift*, where only the outcome distribution changes—also termed *choice-based sampling* or *endogenous stratified sampling* in Manski and Lerman (1977), *prior probability shift* in Storkey (2013), and *target shift* in Scott (2019), Zhang et al. (2013). Another popular condition is *concept shift*, where the covariate or label distribution does not change—also termed *conditional shift* in Zhang et al. (2013). See, for example, Kouw and Loog (2018), Moreno-Torres et al. (2012), Schölkopf et al. (2012), for reviews of common dataset shift conditions. These three conditions—covariate, label, and concept shift—are popular because they are interpretable, broadly applicable, and analytically tractable. There is extensive literature on machine learning under these dataset shift conditions (e.g., Sugiyama, Krauledat and Muller (2007), Lipton, Wang and Smola (2018), Pathak, Ma and Wainwright (2022), Ma, Pathak and Wainwright (2023), among others).

Other conditions and methods have also been studied for more specific problems; examples include (generalized, penalized) linear models (Bastani (2021), Cai, Li and Liu (2022), Chakrabortty and Cai (2018), Gu, Han and Duan (2022), Liu, Zhang and Cai (2020), Liu et al. (2023), Tian and Feng (2022), Zhang, Brown and Cai (2019), Zhang et al. (2022), Zhou et al. (2022)), binary classification (Cai and Wei (2021), Scott (2019)), graphical models (He et al. (2022), Li, Cai and Li (2022)), and location-scale shifts (Zhang et al. (2013)), among others.

For domain adaptation where a limited amount of fully observed data from the target population is available, there may be multiple valid methods to incorporate source population data. It is thus important to understand which ones efficiently extract information from data in both source and target populations. However, the efficient use of source population data to supplement the target population data has only been recently studied. Azriel et al. (2021), Gronsbell et al. (2022), Yuval and Rosset (2023), Zhang, Chakrabortty and Bradic (2021), Zhang and Bradic (2022) studied this problem for mean estimation and (generalized) linear models, under concept shift (i.e., for semisupervised learning). Li and Luedtke (2023) studied efficiency theory for data fusion with an emphasis on causal inference applications, a setting related to ours with a somewhat different primary objective. Other related works in data fusion include Angrist and Krueger (1992), Chatterjee et al. (2016), Chen and Chen (2000), D'Orazio, Di Zio and Scanu (2006, 2010), Evans et al. (2021), Rässler (2012), Robins, Hsieh and Newey (1995), among others. A study of domain adaptation with more general prediction techniques under general dataset shift conditions is lacking.

1.2. *Our contributions.* In this paper, we study the general problem of efficient model-agnostic risk estimation in a target population for data adaptation with fully observed data from both source and target populations under various dataset shift conditions. We take the perspective of modern semiparametric efficiency theory (see, e.g., Bolthausen, Perkins and van der Vaart (2002), Pfanzagl (1985, 1990), van der Vaart (1998)), because many dataset shift conditions can be formulated as *restrictions on the observed data generating mechanism*, yielding a semiparametric model.

We estimate the risk due to its broad applicability and central role in training predictive models and model selection (e.g., Vapnik (1992), Györfi et al. (2002), etc). Empirical risk minimization (ERM) is a fundamental approach in statistics and machine learning, where the goal is to minimize the empirical average of the loss—that is, the risk—over a set of candidate models. Recent studies have highlighted the importance of accurate risk estimation for effective model selection in settings where the risk cannot be estimated nonparametrically due to the presence of high-dimensional nuisance functions (e.g., van der Vaart, Dudoit and Laan (2006), Brookhart and Van Der Laan (2006), Nie and Wager (2021), Foster and Syrgkanis (2023), etc). Our risk estimators have the potential to be relevant and useful in such settings, by enabling improved model selection through improved risk estimation.

Likewise, the related goal of constructing prediction sets with guaranteed coverage (Vovk (2013), Qiu, Dobriban and Tchetgen Tchetgen (2022), Yang, Kuchibhotla and Tchetgen Tchetgen (2022)) often depends on precise estimates of the coverage error probability of constructed prediction sets (Angelopoulos et al. (2021), Park et al. (2020, 2022), Yang, Kuchibhotla and Tchetgen Tchetgen (2022)).

After presenting the general problem setup in Section 2, we consider a general dataset shift condition, which we call *sequential conditionals* (Condition DS.0† in Section 3), for scenarios where target population data is available while the source and target populations may have only partially overlapping support. This condition includes covariate, label and concept shift as special cases. We consider data where an observation Z can be decomposed into components $(Z_1, \dots, Z_K)$. Under this condition, some of the conditional distributions $Z_k \mid (Z_1, \dots, Z_{k-1})$ are shared between the target and source populations, for $k = 1, \dots, K$. As our first main contribution, we propose a novel risk estimator that we formally show in Theorem 1 to be semiparametrically efficient and multiply robust (Tchetgen Tchetgen (2009), Vansteelandt, Rotnitzky and Robins (2007)) under this dataset shift condition.

In particular, we propose to obtain flexible estimators $\hat{\theta}^k$ of certain nuisance conditional odds functions θ_*^k conditional on variables $(Z_1, \dots, Z_k)$, and estimators $\hat{\ell}^k$ of conditional mean loss functions ℓ_*^k by *sequential regression* of $\hat{\ell}^{k+1}(Z_1, \dots, Z_{k+1})$ on $(Z_1, \dots, Z_k)$. We show that our risk estimator is efficient given sufficient convergence rates of the product of errors of (i) $\hat{\theta}^k$ for θ_*^k, and of (ii) $\hat{\ell}^k$ for ℓ_*^k. Moreover, when $\hat{\ell}_v^k$ converges to a certain limit function ℓ_∞^k, our estimator is 2^{K-1}-robust. Specifically, it is consistent if for every $k = 1, \dots, K-1$, $\hat{\theta}^k$ is consistent for θ_*^k *or* $\hat{\ell}^k$ is consistent for the *oracle regression* u^k of ℓ_∞^k; but not necessarily both. The latter oracle regression is defined as the conditional expectation of ℓ_∞^k on $(Z_1, \dots, Z_k)$ under the true distribution.

Our choice of parametrization and the sequential construction of $\hat{\ell}^k$ are key to multiple robustness. Suppose instead that each true conditional mean loss function ℓ_*^k is instead parameterized *directly*—rather than sequentially—as the regression of the loss on the variables $(Z_1, \dots, Z_k)$ in the target population, and accordingly construct $\hat{\ell}^k$ by direct regression in the target population. Then the resulting estimating equation-based estimator using the efficient influence function is not guaranteed to be 2^{K-1}-robust.

Based on this estimator, we further propose a straightforward specification test (Hausman (1978)) of whether our efficient estimator converges to the risk of interest in probability, which can be used to test the assumed sequential conditionals condition. In doing so, we theoretically analyze the behavior of our proposed estimator when the sequential conditionals condition fails. We analytically derive the bias due to the failure of sequential conditionals and show that, in this case, our estimator may diverge arbitrarily as sample size increases if the support of the source populations only partially overlaps with the target population. Under the sequential conditionals condition, such a scenario for the supports is allowed, but does not lead to this convergence issue. To obtain this result, we need a more careful analysis than the standard analysis of Z-estimators (e.g., Section 3.3 in van der Vaart and Wellner (1996)) because of the inconsistency of our estimator without the sequential conditionals condition.

Next, we investigate the efficient risk estimation problem in more detail for concept shift in the features and for covariate shift, in Sections 4 and 5, respectively. We characterize when efficiency gains are large, develop simplified efficient and robust estimators, and study their empirical performance in simulation studies. In particular, we show that our estimator is regular and asymptotically linear (RAL) *even if the nuisance function is estimated inconsistently* under concept shift (Theorem 2). We also show a new impossibility result about such full robustness for covariate or label shift under common parametrizations (Lemma 1).

We present additional new results in the Supplementary Material (Qiu, Tchetgen Tchetgen and Dobriban (2024)). We present additional simulation results showcasing the efficiency gain from our proposed methods in model comparison and model training in Supplement S4. We illustrate our proposed estimators in an HIV risk prediction example in Supplement S5. In Supplement S6, we derive efficiency bounds for risk estimation under three other widely-applicable dataset shift conditions, *posterior drift* (Scott (2019)) *location-scale shift* (Zhang et al. (2013)), and *invariant density ratio shape* motivated by Tasche (2017). The proof of these results requires delicate derivations involving tangent spaces and their orthogonal complements, leading to intricate linear integral equations and, in some cases, a closed-form solution for the efficient influence functions. In Supplement S7, we present additional results on other widely-applicable dataset shift conditions, including the invariant density ratio condition (Tasche (2017)) and stronger versions of posterior drift (Scott (2019)) and location-scale shift (Zhang et al. (2013)) conditions. In Supplement S8, we describe how our proposed risk estimators can help construct prediction sets with marginal or training-set conditional validity. Proofs of all theoretical results can be found in Supplement S9. We implement our proposed methods for covariate, label and concept shift in an R package available at https://github.com/QIU-Hongxiang-David/RiskEstDShift.

2. Problem setup. Let O be a prototypical data point consisting of the observed data Z lying in a space $\mathcal{Z}$ and an integer indexing variable A in a finite set $\mathcal{A}$ containing zero. The variable A indicates whether the data point comes from the target population ($A = 0$) or a source population ($A \in \mathcal{A} \setminus \{0\}$).[1] The observed data $(O_1, \dots, O_n)$ is an independent and identically distributed (i.i.d.) sample from an unknown distribution P_*. We will use a subscript $*$ to denote components of P_* throughout this paper. Data Z is observed from both the source and target populations, that is, for both $A = 0$ and $A \neq 0$.

The estimand of interest is the risk, namely the average value of a given loss function $\ell : \mathcal{Z} \to \mathbb{R}$, in the target population:

$$r_* := R(P_*) := \mathbb{E}_{P_*}\big[\ell(Z) \mid A = 0\big]. \tag{1}$$

We often focus on the supervised setting where $Z = (X, Y)$, with $X \in \mathcal{X}$ being the covariate or feature and $Y \in \mathcal{Y}$ being the outcome or label. In this case, our observed data are i.i.d. triples $(X_i, Y_i, A_i) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{A}$ distributed according to P_*. Next, we provide two examples of loss functions ℓ below.

EXAMPLE 1 (Supervised learning/regression). Let $f : \mathcal{X} \to \mathcal{Y}'$ be a given predictor—obtained, for example, from a separate training dataset—for some space $\mathcal{Y}'$ that can differ from $\mathcal{Y}$. It may be of interest to estimate a measure of the accuracy of f. One common measure is the mean squared error, which is induced by the squared error loss $\ell(x, y) = (y - f(x))^2$. For binary outcomes where $\mathcal{Y} = \{0, 1\}$, it is also common to consider the risk induced by the cross-entropy loss, namely Bernoulli negative log-likelihood $\ell(x, y) = -y \log\{f(x)\} - (1 - y) \log\{1 - f(x)\}$, when $\mathcal{Y}'$ is the unit interval $(0, 1)$ and f outputs a predicted probability. Another common measure of risk is $P_*(Y \neq f(X) \mid A = 0)$, measuring prediction inaccuracy. This is induced by the loss $\ell(x, y) = \mathbb{1}(y \neq f(x))$ when $\mathcal{Y}' = \{0, 1\}$ and f outputs a predicted label.

EXAMPLE 2 (Prediction sets with coverage guarantees). It is often of interest to construct prediction sets with a coverage guarantee. Two popular guarantees are marginal coverage and

[1] Throughout this paper, we emphasize certain aspects of the observed data-generating mechanism when employing the terms “domain” or “population.” For example, when a random sample is drawn from a superpopulation and variables are measured differently for two subsamples, we may treat this data as two samples from two different populations because of the different observed data-generating mechanisms.

training-set conditional—or *probably approximately correct* (PAC)—coverage (Vovk (2013), Park et al. (2020)). To achieve such coverage guarantees, one may estimate—or obtain a confidence interval for—the coverage error of a given prediction set (Vovk (2013), Angelopoulos et al. (2021), Yang, Kuchibhotla and Tchetgen Tchetgen (2022)). Let $C : \mathcal{X} \to 2^{\mathcal{Y}}$ be a given prediction set. With the indicator loss $\ell(x, y) = \mathbb{1}(y \notin C(x))$, the associated risk is the coverage error probability $P_*(Y \notin C(X) \mid A = 0)$ of C in the target population.

REMARK 1 (Broader interpretation of risk estimation problem). Our results in this paper apply to a broader range of problems beyond risk estimation, if those can be mapped to our setup. The loss function ℓ may be interpreted in a broad sense. We list a few examples below. Moreover, the data point Z does not necessarily have to consist of a covariate vector X and an outcome Y. If additional variables W related to Y are observed—for example, outcomes other than Y—these can be leveraged for risk estimation, even if the prediction model only uses the covariate X.

EXAMPLE 3 (Mean estimation). If the estimand of interest is the mean $\mathbb{E}_{P_*}[Z \mid A = 0]$ for $Z \in \mathbb{R}$, we may take ℓ to be the identity function.

EXAMPLE 4 (Quantile estimation). To estimate a quantile of $Z \mid A = 0$, we may consider ℓ ranging over the function class $\{z \mapsto \mathbb{1}(z \le t) : t \in \mathbb{R}\}$.

EXAMPLE 5 (Model comparison). To compare the performance of two methods in the target population, we may take ℓ to be the difference between the loss of these two methods. For example, with $f^{(1)}$ and $f^{(2)}$ denoting two given predictors, we may take $\ell : (x, y) \mapsto (y - f^{(1)}(x))^2 - (y - f^{(2)}(x))^2$ to be the loss difference.

EXAMPLE 6 (Estimating equation). Suppose that the estimand is the solution β_* to an estimating equation $\mathbb{E}_{P_*}[\Psi_\beta(Z) \mid A = 0] = 0$ in β, which includes linear regression, logistic regression, and parametric regression models as special cases. We may consider ℓ ranging over the function class $\{z \mapsto \Psi_\beta(z)\}$ indexed by β and $\hat{r}_\beta$ an estimator of $\mathbb{E}_{P_*}[\Psi_\beta(Z) \mid A = 0]$. Let the estimator $\hat{\beta}$ of β_* be the solution to $\hat{r}_\beta = 0$ in β. Theorem 3.3.1 of van der Vaart and Wellner (1996) implies that reducing the asymptotic variance (as $n \to \infty$) of $\hat{r}$ leads to a reduced asymptotic variance of $\hat{\beta}$.

Without additional conditions[2] on the true data distribution P_*—under a nonparametric model—the source populations are noninformative about the target population because they may differ arbitrarily. In this case, a viable estimator of r_* is *the nonparametric estimator*, the sample mean over the target population data:[3]

$$\hat{r}_{\text{np}} := \frac{\sum_{i=1}^n \mathbb{1}(A_i = 0)\ell(Z_i)}{\sum_{i=1}^n \mathbb{1}(A_i = 0)}. \tag{2}$$

For any scalars $\rho \in (0, 1)$ and $r \in \mathbb{R}$, we define

$$D_{\text{np}}(\rho, r) : o = (z, a) \mapsto \frac{\mathbb{1}(a = 0)}{\rho}\{\ell(z) - r\}. \tag{3}$$

[2] We suppose that $\text{var}_{P_*}(\ell(X, Y) \mid A = 0) < \infty$, a very mild condition, throughout this paper.

[3] In this paper, we define $0/0 = 0$.

We denote $\rho_* := P_*(A = 0) \in (0, 1)$, the true proportion of data from the target population. It is not hard to show that D_{np} is the influence function of $\hat{r}_{\mathrm{np}}$; $\hat{r}_{\mathrm{np}}$ is asymptotically semiparametrically efficient under a nonparametric model and $\sqrt{n}(\hat{r}_{\mathrm{np}} - r_*) \xrightarrow{d} \mathrm{N}(0, \sigma^2_{*,\mathrm{np}})$ with $\sigma^2_{*,\mathrm{np}} := \mathbb{E}_{P_*}[D_{\mathrm{np}}(\rho_*, r_*)(O)^2]$; see Supplement S9.2.

The nonparametric estimator $\hat{r}_{\mathrm{np}}$ ignores data from the source population. If limited data from the target population is available, namely $P_*(A = 0)$ is small, this estimator might not be accurate. This motivates using source population data and plausible conditions to obtain more accurate estimators.

Notation and terminology. We next introduce some notation and terminology. We will use the terms "covariate" and "feature" interchangeably, and similarly for "label" and "outcome." For any nonnegative integers M and N, we use $[M : N]$ to denote the index set $\{M, M + 1, \dots, N\}$ if $M \leq N$ and the empty index set otherwise; we use $[M]$ as a shorthand for $[1 : M]$. For any finite set S, we use $|S|$ to denote its cardinality.

For a distribution P, we use $P_\sharp$ and $P_{\sharp|\natural}$ to denote the marginal distribution of the random variable $\sharp$ and the conditional distribution of $\sharp \mid \natural$, respectively, under P; we use $P_{*,\sharp}$ and $P_{*,\sharp|\natural}$ to denote these distributions under P_*. We use P^n to denote the empirical distribution of a sample of size n from P. When splitting the sample into $V > 0$ folds, we use $P^{n,v}$ and $P^{n,v}_\sharp$ to denote empirical distributions in fold $v \in [V]$. All functions considered will be measurable with respect to appropriate sigma-algebras, which will be kept implicit. For any function f, any distribution P, we sometimes use Pf to denote $\int f \, \mathrm{d}P$. For any $p \in [1, \infty]$, we use $\|f\|_{L^p(P)}$ to denote the $L^p(P)$ norm of f, namely $(\int f(x)^p P(\mathrm{d}x))^{1/p}$. We also use $L^p(P)$ to denote the space of all functions with a finite $L^p(P)$ norm, and use $L^p_0(P)$ to denote $\{f \in L^p(P) : \int f \, \mathrm{d}P = 0\}$. All asymptotic results are with respect to the sample size n tending to infinity.

We finally review a few concepts and basic results that are central to semiparametric efficiency theory. More thorough introductions can be found in Bickel et al. (1993), Bolthausen, Perkins and van der Vaart (2002), Pfanzagl (1985, 1990), van der Vaart (1998). An estimator $\hat{\theta}$ of a parameter $\theta_* = \theta_*(P_*)$ is said to be asymptotically linear if $\hat{\theta} = \theta_* + n^{-1} \sum_{i=1}^n \mathrm{IF}(O_i) + \mathrm{o}_p(n^{-1/2})$ for a function $\mathrm{IF} \in L^2_0(P_*)$. This asymptotic linearity implies that $\sqrt{n}(\hat{\theta} - \theta_*) \xrightarrow{d} \mathrm{N}(0, \mathbb{E}_{P_*}[\mathrm{IF}(O)^2])$. The function IF is called the influence function of $\hat{\theta}$. Under a semiparametric model, there may be infinitely many influence functions, but there exists a unique efficient influence function, which is the influence function of regular asymptotically linear (RAL) estimators with the smallest asymptotic variance. Under a nonparametric model, all RAL estimators of a parameter θ_* share the same influence function, which equals the efficient influence function.

3. Cross-fit estimation under a general dataset shift condition.

3.1. *Statement of condition and efficiency bound.* We consider the following general dataset shift condition characterized by sequentially identical conditional distributions introduced by Li and Luedtke (2023). Under this condition, some auxiliary source population datasets are informative about one component of the target population. In this section, we may use Q to denote the target population and allow data from Q not to be observed, namely $\mathcal{A}$ might not contain index 0. We still use $r_* = \mathbb{E}_Q[\ell(Z)]$ to denote the target population risk. We allow Z to be a general random variable rather than just (X, Y) and allow more than one source population to be present. Thus, let Z be decomposed into $K \geq 1$ components $(Z_1, \dots, Z_K)$. Define $\bar{Z}_0 := \varnothing$, $\bar{Z}_k := (Z_1, \dots, Z_k)$, and $\mathcal{Z}_{k-1}$ to be the support of $\bar{Z}_{k-1} \mid A = 0$ for $k \in [K]$. The condition is as follows.

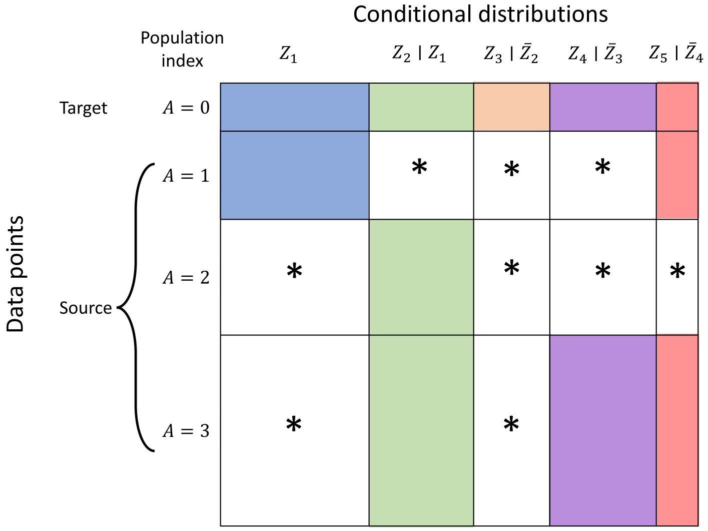

FIG. 1. *Illustration of Condition* DS.0† *with* $K = 5$. *In each column, cells sharing the same color represent the same conditional distribution, while cells with asterisks represent conditional distributions that may arbitrarily differ from the target population* ($A = 0$). *In this example*, $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \{2, 3\}$, $\mathcal{S}_3 = \varnothing$, $\mathcal{S}_4 = \{3\}$, *and* $\mathcal{S}_5 = \{1, 3\}$.

CONDITION DS.0 (General sequential conditionals). For every $k \in [K]$, there exists a known nonempty subset $\mathcal{S}'_k \subset \mathcal{A}$ such that, (i) the distribution of $\bar{Z}_{k-1}$ under Q is dominated by $\bar{Z}_{k-1} \mid A \in \mathcal{S}'_k$ under P_*, and (ii) for all $a \in \mathcal{S}'_k$, $Z_k \mid \bar{Z}_{k-1} = \bar{z}_{k-1}, A = a$ is distributed identically to $Z_k \mid \bar{Z}_{k-1} = \bar{z}_{k-1}$ under Q for Q-almost every $\bar{z}_{k-1}$ in the support $\mathcal{Z}_{k-1}$ of $\bar{Z}_{k-1}$ under Q.

This condition states that conditionally on $A \in \mathcal{S}'_k$, Z_k is independent of A given $\bar{Z}_{k-1}$. Equivalently, it states that every conditional distribution $Z_k \mid \bar{Z}_{k-1}$ ($k \in [K]$) under Q is equal to that in the source populations with $a \in \mathcal{S}'_k$. One important special case is when data from the target population is also observed and data from source populations are used to improve efficiency, as stated in the following condition.

CONDITION DS.0† (Sequential conditionals). Condition DS.0 holds with $0 \in \mathcal{A}$ and $0 \in \mathcal{S}'_k$ for every k.

The condition $0 \in \mathcal{S}'_k$ is purely a matter of notation since 0 is the index for the target population Q. The dominance of the distribution of $\bar{Z}_{k-1}$ under Q by the source population in Condition DS.0(i) is automatically satisfied since $0 \in \mathcal{S}'_k$. When working under this stronger condition, we use $\mathcal{S}_k$ to denote $\mathcal{S}'_k \setminus \{0\}$ for short. We show an example of this condition in Figure 1. In particular, we allow for cases where no source population exists to supplement learning some conditional distributions, that is, $\mathcal{S}_k$ may be empty for some k. We also allow irrelevant variables in some source populations to be missing; for example, in Figure 1, (Z_3, Z_4, Z_5) in the source population $A = 2$ may be missing as these variables are not assumed to be informative about the target population.

According to the well-known review by Moreno-Torres et al. (2012), the following four dataset shift conditions are among the most widely considered when one source population is available, so that $\mathcal{A} = \{0, 1\}$. These conditions are all special cases of Condition DS.0†.

CONDITION DS.1 (Concept shift in the features). $X \perp\!\!\!\perp A$.

CONDITION DS.2 (Concept shift in the labels). $Y \perp\!\!\!\perp A$.

CONDITION DS.3 (Full-data covariate shift). $Y \perp\!\!\!\perp A \mid X$.

CONDITION DS.4 (Full-data label shift). $X \perp\!\!\!\perp A \mid Y$.

Condition DS.0† reduces to DS.1—concept shift in the features—by setting $K = 2$, $\mathcal{A} = \{0, 1\}$, $\mathcal{S}_1 = \{1\}$, $\mathcal{S}_2 = \varnothing$, $Z_1 = X$, and $Z_2 = Y$. Indeed, Condition DS.0† for $k = 1$ states that $(Z_1|A = 1) =_d (Z_1|A = 0)$, or equivalently that $(X|A = 1) =_d (X|A = 0)$, which means that $X \perp\!\!\!\perp A$. Since $\mathcal{S}_2 = \varnothing$, Condition DS.0† for $k = 2$ does not impose additional constraints. Similarly, Condition DS.0† reduces to DS.2—concept shift in the labels—with the above choices but switching $Z_1 = Y$ and $Z_2 = X$.

Condition DS.0† reduces to DS.3—full-data covariate shift—by setting $K = 2$, $\mathcal{A} = \{0, 1\}$, $\mathcal{S}_1 = \varnothing$, $\mathcal{S}_2 = \{1\}$, $Z_1 = X$, and $Z_2 = Y$. Indeed, since $\mathcal{S}_1 = \varnothing$, Condition DS.0† for $k = 1$ does not impose constraints. For $k = 2$, Condition DS.0† states that $(Z_2|Z_1, A = 1) =_d (Z_2|Z_1, A = 0)$, or equivalently that $(Y|X, A = 1) =_d (Y|X, A = 0)$, which means that $Y \perp\!\!\!\perp A \mid X$. Similarly, it reduces to DS.4—full-data label shift—with the above choices but switching $Z_1 = Y$ and $Z_2 = X$. We refer to Conditions DS.3 and DS.4 as *full-data* covariate and label shift, respectively, to emphasize that we have full observations (X, Y, A) from the target population. For brevity, we refer to them as covariate or label shift when no confusion shall arise. Condition DS.0† also includes more sophisticated dataset shift conditions and we provide a few examples in Supplement S1 for further examples.

Compared to Condition DS.0†, the more general condition DS.0 may also be applicable to cases without observing data from the target population, for example, covariate shift with unlabeled target population data and labeled source population data.

In our problem, a plug-in estimation approach could be to pool all data sets that share the same conditional distributions for each k, namely $A \in \mathcal{S}'_k$, and estimating these distributions nonparametrically. Then, the risk can be estimated by integrating the loss over the estimated distribution. However, as it is well known, this approach can suffer from a large bias or may limit the choice of distribution or density estimators, and typically requires a delicate case-by-case analysis to establish its accuracy (e.g., McGrath and Mukherjee (2022), etc.).

We now describe and review results on efficiency, which characterize the smallest possible asymptotic variance of a sequence of regular estimators under the dataset shift condition DS.0. This will form the basis of our proposed estimator in the next section. We first introduce a few definitions. For Condition DS.0, let λ_*^{k-1} denote the Radon-Nikodym derivative of the distribution of $\bar{Z}_{k-1}$ under Q relative to that of $\bar{Z}_{k-1} \mid A \in \mathcal{S}'_k$ under P_*. For Condition DS.0†, define the conditional probability of each population

$$\Pi_*^{k,a} : \bar{z}_k \mapsto P_*(A = a \mid \bar{Z}_k = \bar{z}_k) \quad \text{for } k \in [0 : K-1],\ a \in \mathcal{A},$$

and let $\pi_*^a := \Pi_*^{0,a} = P_*(A = a)$ denote the marginal probabilities of all populations ($a \in \mathcal{A}$); thus, $\pi_*^0 = \rho_*$ for ρ_* from Section 2. Define the conditional odds of relevant source populations versus the target population: $\theta_*^{k-1} := \sum_{a \in \mathcal{S}_k} \Pi_*^{k-1,a} / \Pi_*^{k-1,0}$ for $k \in [K]$. Under the stronger condition DS.0†, it follows from Bayes' theorem that

$$\lambda_*^{k-1} = \frac{\sum_{a \in \mathcal{S}'_k} \pi_*^a}{\pi_*^0 (1 + \theta_*^{k-1})}. \tag{4}$$

For both Conditions DS.0 and DS.0†, we also define conditional means of the loss starting with $\ell_*^K := \ell$ and letting recursively

$$\ell_*^k : \bar{z}_k \mapsto \mathbb{E}_{P_*}\big[\ell_*^{k+1}(\bar{Z}_{k+1}) \mid \bar{Z}_k = \bar{z}_k, A \in \mathcal{S}'_{k+1}\big] \quad \text{for } \bar{z}_k \in \mathcal{Z}_k,\ k \in [K-1]. \tag{5}$$

We allow ℓ_*^k to take any value outside $\mathcal{Z}_k$, for example, when the support of $\bar{Z}_k \mid A \in \mathcal{S}'_{k+1}$ is larger than the support $\mathcal{Z}_k$ of $\bar{Z}_k \mid A = 0$. Under Condition DS.0, $\ell_*^k(\bar{z}_k) = \mathbb{E}_Q[\ell(Z) \mid \bar{Z}_k = \bar{z}_k]$ for $\bar{z}_k \in \mathcal{Z}_k$. We discuss the consequences of the nonunique definition of ℓ_*^k without Condition DS.0 in more detail in Section 3.4. Let $\boldsymbol{\ell}_* := (\ell_*^k)_{k=1}^{K-1}$, $\boldsymbol{\lambda}_* := (\lambda_*^k)_{k=1}^{K-1}$, $\boldsymbol{\theta}_* := (\theta_*^k)_{k=1}^{K-1}$, and $\boldsymbol{\pi}_* := (\pi_*^a)_{a \in \mathcal{A}}$ be collections of true nuisances. For any given collections $\boldsymbol{\ell} = (\ell^k)_{k=1}^{K-1}$, $\boldsymbol{\lambda} := (\lambda^k)_{k=1}^{K-1}$, $\boldsymbol{\theta} = (\theta^k)_{k=1}^{K-1}$ and $\boldsymbol{\pi} := (\pi^a)_{a \in \mathcal{A}}$ of nuisances, a scalar r and $\ell^K := \ell$, define the pseudo-losses $\widetilde{\mathcal{T}}(\boldsymbol{\ell}, \boldsymbol{\lambda}, \boldsymbol{\pi})$ and $\mathcal{T}(\boldsymbol{\ell}, \boldsymbol{\theta}, \boldsymbol{\pi}) : \mathcal{O} \to \mathbb{R}$ based on these nuisances, so that for $o = (z, a)$,[4]

$$\widetilde{\mathcal{T}}(\boldsymbol{\ell}, \boldsymbol{\lambda}, \boldsymbol{\pi})(o) = \sum_{k=2}^{K} \frac{\mathbb{1}(a \in \mathcal{S}'_k)}{\sum_{b \in \mathcal{S}'_k} \pi^b} \lambda^{k-1}(\bar{z}_{k-1})\big[\ell^k(\bar{z}_k) - \ell^{k-1}(\bar{z}_{k-1})\big] + \frac{\mathbb{1}(a \in \mathcal{S}'_1)}{\sum_{b \in \mathcal{S}'_1} \pi^b} \ell^1(z_1), \tag{6}$$

$$\mathcal{T}(\boldsymbol{\ell}, \boldsymbol{\theta}, \boldsymbol{\pi})(o) = \sum_{k=2}^{K} \frac{\mathbb{1}(a \in \mathcal{S}'_k)}{\pi^0(1 + \theta^{k-1}(\bar{z}_{k-1}))} \{\ell^k(\bar{z}_k) - \ell^{k-1}(\bar{z}_{k-1})\} + \frac{\mathbb{1}(a \in \mathcal{S}'_1)}{\pi^0(1 + \theta^0)} \ell^1(z_1). \tag{7}$$

The motivation for this transformation is similar to that for the unbiased transformation from Rotnitzky, Faraggi and Schisterman (2006), Rubin and van der Laan (2007) and the pseudo-outcome from Kennedy (2020). Further, given any scalar r, with $\theta^0 := \sum_{a \in \mathcal{S}_1} \pi^a / \pi^0$, define

$$D_{\mathrm{GSC}}(\boldsymbol{\ell}, \boldsymbol{\lambda}, \boldsymbol{\pi}, r) : o = (z, a) \mapsto \widetilde{\mathcal{T}}(\boldsymbol{\ell}, \boldsymbol{\lambda}, \boldsymbol{\pi})(o) - \frac{\mathbb{1}(a \in \mathcal{S}'_1)}{\sum_{b \in \mathcal{S}'_1} \pi^b} r, \tag{8}$$

$$D_{\mathrm{SC}}(\boldsymbol{\ell}, \boldsymbol{\theta}, \boldsymbol{\pi}, r) : o = (z, a) \mapsto \mathcal{T}(\boldsymbol{\ell}, \boldsymbol{\theta}, \boldsymbol{\pi})(o) - \frac{\mathbb{1}(a \in \mathcal{S}'_1)}{\pi^0(1 + \theta^0)} r. \tag{9}$$

A key result we will use is that the efficient influence function for estimating r_* (i) equals $D_{\mathrm{GSC}}(\boldsymbol{\ell}_*, \boldsymbol{\lambda}_*, \boldsymbol{\pi}_*, r_*)$ under Condition DS.0 and when λ_*^k are uniformly bounded away from zero and infinity as a function of $\bar{z}_k \in \mathcal{Z}_k$ for all $k \in [0 : K-1]$, and (ii) specializes to $D_{\mathrm{SC}}(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*, \boldsymbol{\pi}_*, r_*)$ under Condition DS.0† and when θ_*^k are bounded functions for all $k \in [0 : K-1]$. This follows by Theorem 2 and Corollary 1 in Li and Luedtke (2023). Consequently, the smallest possible asymptotic variance of a sequence of $n^{1/2}$-scaled RAL estimators is

$$\sigma^2_{*,\mathrm{GSC}} := \mathbb{E}_{P_*}\big[D_{\mathrm{GSC}}(\boldsymbol{\ell}_*, \boldsymbol{\lambda}_*, \boldsymbol{\pi}_*, r_*)(O)^2\big] \tag{10}$$

under Condition DS.0 and specializes to $\sigma^2_{*,\mathrm{SC}} := \mathbb{E}_{P_*}[D_{\mathrm{SC}}(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*, \boldsymbol{\pi}_*, r_*)(O)^2]$ under Condition DS.0†. Despite the possible nonunique definition of conditional mean loss $\boldsymbol{\ell}_*$, both $D_{\mathrm{GSC}}(\boldsymbol{\ell}_*, \boldsymbol{\lambda}_*, \boldsymbol{\pi}_*, r_*)$ and $D_{\mathrm{SC}}(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*, \boldsymbol{\pi}_*, r_*)$ are uniquely defined under Condition DS.0 and DS.0†, respectively. Here, we have used the odds parametrization rather than the density ratio or Radon-Nikodym derivative parametrization from Li and Luedtke (2023) for Condition DS.0† because the former is often more convenient for estimation.

3.2. *Cross-fit risk estimator.*

We next present our proposed estimator along with the motivation. All estimators will implicitly depend on the sample size n, but we will sometimes omit this dependence from notation for conciseness. We take as given a flexible regression method $\mathcal{K}$ estimating conditional means and a flexible classifier $\mathcal{C}$ estimating conditional odds—both taking outcome, covariates, and an index set for data points being used as inputs in order. For example, $\mathcal{K}$ and $\mathcal{C}$ may be random forests, neural networks, gradient boosting, or an ensemble learner. We also take as given a flexible density ratio estimator $\mathcal{W}$, taking an

[4] If $\theta^{k-1}(\bar{z}_{k-1}) = \infty$, we set $1/(1 + \theta^{k-1}(\bar{z}_{k-1}))$ to be zero. When θ^{k-1} equals the truth θ_*^{k-1}, this case can happen for $\bar{z}_{k-1}$ outside the support of $\bar{Z}_{k-1} \mid A = 0$ but inside the support of $\bar{Z}_{k-1} \mid A \in \mathcal{S}_k$.

Algorithm 1† Cross-fit estimator of $r_* = \mathbb{E}_{P_*}[\ell(Z) \mid A = 0]$ under Condition DS.0†

```
Require: Data $\{O_i = (Z_i, A_i)\}_{i=1}^n$, relevant source population sets $\mathcal{S}'_k$ ($k \in [K]$), number $V$ of folds, classifier $\mathcal{C}$, regression estimator $\mathcal{K}$
Randomly split data into $V$ folds of approximately equal sizes. Denote the index set of data points in fold $v$ by $I_v$, and the index set of data points with $A \in \mathcal{S}'_k$ by $J_k$ for $k \in [K]$.
for $v \in [V]$ do
    For all $k \in [K-1]$, estimate $\theta_*^k$ using data out of fold $v$, by classifying $A = 0$ against $A \in \mathcal{S}_{k+1}$ via the classifier $\mathcal{C}$ with covariates $\bar{Z}_k$ in the subsample with $A \in \mathcal{S}'_{k+1}$; that is, set $\hat{\theta}_v^k := \mathcal{C}(\mathbb{1}(A \neq 0), \bar{Z}_k, ([n] \setminus I_v) \cap J_{k+1})$ and $\widehat{\boldsymbol{\theta}}_v := (\hat{\theta}_v^k)_{k=1}^{K-1}$.
    Set $\hat{\pi}_v^a := |I_v|^{-1} \sum_{i \in I_v} \mathbb{1}(A_i = a)$ for all $a \in \mathcal{A}$, $\widehat{\boldsymbol{\pi}}_v := (\hat{\pi}_v^a)_{a \in \mathcal{A}}$, $\hat{\theta}_v^0 := \sum_{a \in \mathcal{S}_1} \hat{\pi}_v^a / \hat{\pi}_v^0$, and $\hat{\ell}_v^K$ to be $\ell$.
    for $k = K-1, \dots, 1$ do
        Estimate $\ell_*^k$ using data out of fold $v$ by regressing $\hat{\ell}_v^{k+1}(\bar{Z}_{k+1})$ on $\bar{Z}_k$ in the subsample with $A \in \mathcal{S}'_{k+1}$; that is, set $\hat{\ell}_v^k := \mathcal{K}(\hat{\ell}_v^{k+1}(\bar{Z}_{k+1}), \bar{Z}_k, ([n] \setminus I_v) \cap J_{k+1})$.
    Set $\widehat{\boldsymbol{\ell}}_v := (\hat{\ell}_v^k)_{k=1}^{K-1}$.
    Compute an estimator of $r_*$ for fold $v$:
    (11)    $\hat{r}_v := \frac{1}{|I_v|} \sum_{i \in I_v} \mathcal{T}(\widehat{\boldsymbol{\ell}}_v, \widehat{\boldsymbol{\theta}}_v, \widehat{\boldsymbol{\pi}}_v)(O_i)$
Compute the cross-fit estimator combining estimators $\hat{r}_v$ from all folds: $\hat{r} := \frac{1}{n} \sum_{v=1}^V |I_v| \hat{r}_v$.
```

index set for data points being used as input, which may be transformed from a classifier $\mathcal{C}$ by Bayes' theorem or based on kernel density estimators.

We take Condition DS.0† as an example to illustrate ideas behind our proposed estimator. One approach to constructing an efficient estimator of r_* is to solve the estimating equation $\sum_{i=1}^n D_{\mathrm{SC}}(\widehat{\boldsymbol{\ell}}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\pi}}, r)(O_i) = 0$ for r, where $\widehat{\boldsymbol{\ell}} = \widehat{\boldsymbol{\ell}}_n$, $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_n$ and $\widehat{\boldsymbol{\pi}} = \widehat{\boldsymbol{\pi}}_n$ are estimators of nuisances $\boldsymbol{\ell}_*$, $\boldsymbol{\theta}_*$ and $\boldsymbol{\pi}_*$, respectively, and use the solution as the estimator. See Section 7.1, Part III in Bolthausen, Perkins and van der Vaart (2002) for a more thorough introduction to achieving efficiency by solving an estimating equation. We further use sample splitting (Hajek (1962), Bickel (1982), Schick (1986), Chernozhukov et al. (2018)) to allow for more flexible estimators, leading to our proposed estimator in Algorithm 1†. Splitting the sample into a fixed number V of folds I_v ($v \in [V]$) leads to the estimating equation $\sum_{i \in I_v} D_{\mathrm{SC}}(\widehat{\boldsymbol{\ell}}_v, \widehat{\boldsymbol{\theta}}_v, \widehat{\boldsymbol{\pi}}_v, r)(O_i) = 0$ averaging over data in each every fold v, with preliminary estimators $(\widehat{\boldsymbol{\ell}}_v, \widehat{\boldsymbol{\theta}}_v, \widehat{\boldsymbol{\pi}}_v)$ using data outside of the fold. The solution is given in (11). The corresponding estimator for more general condition DS.0 is similar and described in Algorithm 1 in the Supplementary Material.

REMARK 2 (Estimation of marginal probabilities $\boldsymbol{\pi}_*$). It is viable to replace the in-fold estimator $\widehat{\boldsymbol{\pi}}_v$ of π_* with an out-of-fold estimator in Algorithm 1†. These two approaches have the same theoretical properties that we will show next, and similar empirical performance.

We next present sufficient conditions for the asymptotic efficiency and multiple robustness of the estimator $\hat{r}$. In the following analyses without assuming Condition DS.0, we assume that nuisance estimators $\widehat{\boldsymbol{\ell}}_v$ can be evaluated at any point in the space $\mathcal{Z}$ containing the observation, even if that point is outside the support under P_*. For illustration, we assume that, for each $k \in [2 : K]$, $\|\hat{\ell}_v^k - \ell_\infty^k\|_{L^2(\nu_k)} \xrightarrow{p} 0$ for some function ℓ_∞^k, where ν_k denotes the distribution of $\bar{Z}_k \mid A \in \mathcal{S}_k'$ under P^0. Define the *oracle estimator* h_v^{k-1} of ℓ_*^{k-1} based on $\hat{\ell}_v^k$, evaluated under the true distribution P_*, as[5]

$$h_v^{k-1} : \bar{z}_{k-1} \mapsto \mathbb{E}_{P_*}[\hat{\ell}_v^k(\bar{Z}_k) \mid \bar{Z}_{k-1} = \bar{z}_{k-1}, A \in \mathcal{S}_k'],$$

and the product bias term $B_{k,v}$ as

$$\begin{aligned} &\frac{\sum_{a \in \mathcal{S}_k'} \pi_*^a}{\sum_{a \in \mathcal{S}_k'} \hat{\pi}_v^a} \mathbb{E}_{P_*}\big[\{\hat{\lambda}_v^{k-1}(\bar{Z}_{k-1}) - \lambda_*^{k-1}(\bar{Z}_{k-1})\} \\ &\quad \times \{h_v^{k-1}(\bar{Z}_{k-1}) - \hat{\ell}_v^{k-1}(\bar{Z}_{k-1})\} \mid A \in \mathcal{S}_k'\big] \\ &\quad + \left\{\frac{\sum_{a \in \mathcal{S}_k'} \pi_*^a}{\sum_{a \in \mathcal{S}_k'} \hat{\pi}_v^a} - \frac{\sum_{a \in \mathcal{S}_1'} \pi_*^a}{\sum_{a \in \mathcal{S}_1'} \hat{\pi}_v^a}\right\} \mathbb{E}_Q\big[h_v^{k-1}(\bar{Z}_{k-1}) - \hat{\ell}_v^{k-1}(\bar{Z}_{k-1})\big] \end{aligned} \tag{12}$$

for Condition DS.0 and Algorithm 1, which reduces to

$$\begin{aligned} &\frac{\sum_{a \in \mathcal{S}_k'} \pi_*^a}{\sum_{a \in \mathcal{S}_k'} \hat{\pi}_v^a} \mathbb{E}_{P_*}\Bigg[\left\{\frac{\sum_{a \in \mathcal{S}_k'} \hat{\pi}_v^a}{\hat{\pi}_v^0(1 + \hat{\theta}_v^{k-1}(\bar{Z}_{k-1}))} - \frac{\sum_{a \in \mathcal{S}_k'} \pi_*^a}{\pi_*^0(1 + \theta_*^{k-1}(\bar{Z}_{k-1}))}\right\} \\ &\quad \times \{h_v^{k-1}(\bar{Z}_{k-1}) - \hat{\ell}_v^{k-1}(\bar{Z}_{k-1})\} \mid A \in \mathcal{S}_k'\Bigg] \\ &\quad + \left(\frac{\sum_{a \in \mathcal{S}_k'} \pi_*^a}{\sum_{a \in \mathcal{S}_k'} \hat{\pi}_v^a} - \frac{\sum_{a \in \mathcal{S}_1'} \pi_*^a}{\sum_{a \in \mathcal{S}_1'} \hat{\pi}_v^a}\right) \mathbb{E}_{P_*}\big[h_v^{k-1}(\bar{Z}_{k-1}) - \hat{\ell}_v^{k-1}(\bar{Z}_{k-1}) \mid A = 0\big] \end{aligned} \tag{13}$$

for Condition DS.0† and Algorithm 1† when $\hat{\lambda}_v^{k-1}$ is transformed from $\widehat{\theta_v^{k-1}}$ and $\hat{\pi}_v$ as in (4).

CONDITION ST.1. For every fold $v \in [V]$,

1. the following term is $\mathrm{o}_p(n^{-1/2})$ for Condition DS.0 and Algorithm 1, or for Condition DS.0† and Algorithm 1†:

$$\sum_{k=2}^{K} B_{k,v}; \tag{14}$$

2. the following term is $\mathrm{o}_p(1)$ for Condition DS.0 and Algorithm 1, or for Condition DS.0† and Algorithm 1†, respectively:

$$\left\| \left(\sum_{a \in \mathcal{S}_1'} \hat{\pi}_v^a\right) \tilde{\mathcal{T}}(\widehat{\boldsymbol{\ell}}_v, \widehat{\boldsymbol{\lambda}}_v, \widehat{\boldsymbol{\pi}}_v) - \left(\sum_{a \in \mathcal{S}_1'} \pi_*^a\right) \tilde{\mathcal{T}}(\boldsymbol{\ell}_*, \boldsymbol{\lambda}_*, \boldsymbol{\pi}_*) \right\|_{L^2(P_*)} \tag{15}$$

$$\text{or} \quad \left\| \left(\sum_{a \in \mathcal{S}_1'} \hat{\pi}_v^a\right) \mathcal{T}(\widehat{\boldsymbol{\ell}}_v, \widehat{\boldsymbol{\theta}}_v, \widehat{\boldsymbol{\pi}}_v) - \left(\sum_{a \in \mathcal{S}_1'} \pi_*^a\right) \mathcal{T}(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*, \boldsymbol{\pi}_*) \right\|_{L^2(P_*)}. \tag{16}$$

In each part of the condition, the requirement for the more general condition DS.0 reduces to that for the more restrictive condition DS.0† with $\hat{\lambda}_v^{k-1} = \sum_{a \in \mathcal{S}_k'} \hat{\pi}_v^a / \{\hat{\pi}_v^0(1 + \hat{\theta}_v^{k-1})\}$. To illustrate Condition ST.1, define the *limiting oracle estimator* u^{k-1} of ℓ_*^{k-1} based on ℓ_∞^k as

$$u^{k-1} : \bar{z}_{k-1} \mapsto \mathbb{E}_{P_*}[\ell_\infty^k(\bar{Z}_k) \mid \bar{Z}_{k-1} = \bar{z}_{k-1}, A \in \mathcal{S}_k'].$$

[5] In all expectations involving nuisance estimators such as $(\widehat{\boldsymbol{\ell}}_v, \widehat{\boldsymbol{\theta}}_v, \widehat{\boldsymbol{\pi}}_v)$, these estimators are treated as fixed and the expectation integrates over the randomness in a data point $O = (Z, A)$. For example, $\mathbb{E}_{P_*}[\hat{\ell}_v^1(Z_1) \mid A = 0] = \int \hat{\ell}_v^1 \, \mathrm{d}\mu$ where μ is the distribution of $Z_1 \mid A = 0$ under P_*, and this expectation is itself random due to the randomness in $\hat{\ell}_v^1$.

By the definitions of h_v^{k-1} and u^{k-1}, we have that $h_v^{k-1}(\bar{Z}_{k-1}) - \hat{\ell}_v^{k-1}(\bar{Z}_{k-1})$ equals

$$\mathbb{E}_{P_*}\big[\hat{\ell}_v^k(\bar{Z}_k) - \ell_\infty^k(\bar{Z}_k) \mid \bar{Z}_{k-1}, A \in \mathcal{S}_k'\big] + u^{k-1}(\bar{Z}_{k-1}) - \hat{\ell}_v^{k-1}(\bar{Z}_{k-1}). \tag{17}$$

Thus, Condition ST.1 would hold if the nuisance estimator $(\widehat{\boldsymbol{\ell}}_v, \widehat{\boldsymbol{\theta}}_v)$ converge to the truth $(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*)$ sufficiently fast. Under Condition DS.0 or DS.0†, part 2 is a consistency condition that is often mild; we discuss the other case in Section 3.4. We next focus on part 1 and consider Condition DS.0† and Algorithm 1† first. The term in (14) is a drift term characterizing the bias of the estimated pseudo-loss $\mathcal{T}(\widehat{\boldsymbol{\ell}}_v, \widehat{\boldsymbol{\theta}}_v, \widehat{\boldsymbol{\pi}}_v)$ due to estimating nuisance functions. Conditions requiring such terms to be $\mathrm{o}_p(n^{-1/2})$ are prevalent in the literature on inference under nonparametric or semiparametric models and are often necessary to achieve efficiency (see, e.g., Newey (1994), Chen and Pouzo (2015), Chernozhukov et al. (2017), Van der Laan and Rose (2018)). Balakrishnan, Kennedy and Wasserman (2023) suggest that such $\mathrm{o}_p(n^{-1/2})$ conditions might be necessary without additional assumptions such as smoothness or sparsity on $\boldsymbol{\ell}_*$ or $\boldsymbol{\theta}_*$. Since $\hat{\pi}_v^a$ is root-n consistent for π_*^a, the second term in B_k is $\mathrm{o}_p(n^{-1/2})$ under the mild consistency condition that all $\hat{\ell}_v^{k-1}$ are consistent for ℓ_*^{k-1} $(k \in [2:K])$.

By Jensen's inequality and the Cauchy-Schwarz inequality, we have that, for each $k \in [2:K]$, the first term in B_k is $\mathrm{o}_p(n^{-1/2})$ if both $\hat{\ell}_v^{k-1} - h_v^{k-1}$ and $1/(1+\hat{\theta}_v^{k-1}) - 1/(1+\theta_*^{k-1})$ converge to zero in probability at rates faster than $n^{-1/4}$. We illustrate this rate requirement for $\hat{\ell}_v^{k-1} - h_v^{k-1}$. With $\hat{\ell}_v^{k-1}$ estimated by highly adaptive lasso with squared error loss, if h_v^{k-1} has finite variation norm and $\hat{\ell}_v^k$ is bounded, then $\hat{\ell}_v^{k-1} - h_v^{k-1}$ diminishes at a rate faster than $n^{-1/4}$ (Benkeser and van der Laan (2016), van der Laan (2017)). We formally present more interpretable sufficient conditions for part 1 of Condition ST.1 along with some examples of regression methods in Supplement S3. We also allow one difference to converge slower as long as the other converges fast enough to compensate. In principle, it is also possible to empirically check whether the magnitude of the term in (14) is sufficiently small under certain conditions by using methods proposed by Liu, Mukherjee and Robins (2020, 2023). We do not pursue this direction in this paper as it is beyond the scope. For Condition DS.0 and Algorithm 1, sufficient conditions for part 1 of Condition ST.1 depend on how the Radon-Nikodym derivatives $\boldsymbol{\lambda}_*$ are estimated. If the estimators $\widehat{\boldsymbol{\lambda}}_v$ are based on kernel density estimators, part 1 would require strong smoothness conditions; on the other hand, if it is based on a classifier as in (4), which is applicable for the special case of covariate shift without labeled target population data, the discussion above still applies.

CONDITION ST.2. For each fold $v \in [V]$, Condition ST.1 holds with the $\mathrm{o}_p(n^{-1/2})$ in part 1 replaced by $\mathrm{o}_p(1)$ and the $\mathrm{o}_p(1)$ in part 2 replaced by $\mathrm{O}_p(1)$.

Condition ST.2 is much weaker than Condition ST.1. We illustrate this for Condition DS.0† and Algorithm 1†. By (17) and the assumption that $\hat{\ell}_v^k$ converges to ℓ_∞^k in probability, Condition ST.2 holds if, for each $k \in [2:K]$, either $1/(1+\hat{\theta}_v^{k-1})$ is consistent for $1/(1+\theta_*^{k-1})$ or $\hat{\ell}_v^{k-1}$ is consistent for u^{k-1}. Thus, for each fold v, there are 2^{K-1} possible ways for some of nuisance function estimators $\hat{\theta}_v^{k-1}$ and $\hat{\ell}_v^{k-1}$ $(k \in [2:K])$ to be inconsistent while Condition ST.2 still holds. The same multiple allowance of inconsistent estimation applies to Condition DS.0 and Algorithm 1.

REMARK 3. Conditions ST.1 and ST.2 can hold even if the limit ℓ_∞^k of the nuisance function estimator $\hat{\ell}_v^k$ does not exist. This case could happen if the support of $\bar{Z}_k \mid A \in \mathcal{S}_k$ is larger than $\mathcal{Z}_k$. Our results allow for such cases; we introduced the limit ℓ_∞^k to illustrate Conditions ST.1 and ST.2.

These two conditions are used in the next result on $\hat{r}$.

THEOREM 1. *With nuisance estimators $\widehat{\boldsymbol{\ell}}_v, \widehat{\boldsymbol{\lambda}}_v, \widehat{\boldsymbol{\theta}}_v$ and $\widehat{\boldsymbol{\pi}}_v$ in Algorithms 1*† *and the corresponding algorithm for the more general condition from the Supplementary Material, define*[6]

$$\Delta_v := \frac{\sum_{a \in \mathcal{S}_1'} \pi_*^a}{\sum_{a \in \mathcal{S}_1'} \hat{\pi}_v^a} \sum_{k=1}^{K} \mathbb{E}_{P_*}\big[h_v^{k-1}(\bar{Z}_{k-1}) - \hat{\ell}_v^k(\bar{Z}_k) \mid A = 0\big] \tag{18}$$

for every fold $v \in [V]$ and $\Delta := n^{-1} \sum_{v=1}^{V} |I_v| \Delta_v$. The following finite-sample expansion of $\hat{r}$ holds: $\hat{r} - \Delta - r_*$ equals

$$\begin{aligned} &\sum_{v \in [V]} \frac{|I_v|}{n \sum_{a \in \mathcal{S}_1'} \hat{\pi}_v^a} (P^{n,v} - P_*) \Big\{ \Big(\sum_{a \in \mathcal{S}_1'} \hat{\pi}_v^a \Big) \widetilde{\mathcal{T}}(\widehat{\boldsymbol{\ell}}_v, \widehat{\boldsymbol{\lambda}}_v, \widehat{\boldsymbol{\pi}}_v) - \Big(\sum_{a \in \mathcal{S}_1'} \pi_*^a \Big) \widetilde{\mathcal{T}}(\boldsymbol{\ell}_*, \boldsymbol{\lambda}_*, \boldsymbol{\pi}_*) \Big\} \\ &+ \sum_{v \in [V]} \frac{|I_v| \sum_{a \in \mathcal{S}_1'} \hat{\pi}_v^a}{n} \sum_{k=2}^{K} B_{k,v} \\ &+ \sum_{v \in [V]} \frac{|I_v| \sum_{a \in \mathcal{S}_1'} \pi_*^a}{n \sum_{a \in \mathcal{S}_1'} \pi_v^a} (P^{n,v} - P_*) D_{\mathrm{GSC}}(\boldsymbol{\ell}_*, \boldsymbol{\lambda}_*, \boldsymbol{\pi}_*, r_*). \end{aligned} \tag{19}$$

Moreover, if for all n, k, v,

1. $\Pr(\|\hat{\lambda}_v^{k-1} - \lambda_*^{k-1}\|_{L^2(P_*)} > a_{n,k,v}) \leq c_{n,k,v}$ *and* $\Pr(\|\hat{\ell}_v^{k-1} - h_v^{k-1}\|_{L^2(P_*)} > b_{n,k,v}) \leq d_{n,k,v}$ *for some positive numbers $a_{n,k,v}$, $b_{n,k,v}$, $c_{n,k,v}$ and $d_{n,k,v}$,*
2. $\hat{\lambda}_v^{k-1}$ *are bounded for all k and v, and*
3. $\mathbb{E}_{P_*} |D_{\mathrm{GSC}}(\boldsymbol{\ell}_*, \boldsymbol{\lambda}_*, \boldsymbol{\pi}_*, r_*)(O)|^3 < \infty$,

then for any $\varepsilon > 0$, there exist quantities $\mathscr{C}_1, \mathscr{C}_2 > 0$ that may only depend on ε, P_* and the bound on $\hat{\lambda}_v^{k-1}$ only such that, for any $t > \mathscr{C}_1 \sum_{v \in [V]} \sum_{k=2}^{K} (a_{n,k,v} b_{n,k,v} + a_{n,k,v} + b_{n,k,v} + n^{-1})$, the following finite-sample confidence guarantee holds: $\Pr(|\hat{r} - \Delta - r_*| > t)$ *is at most*

$$\begin{aligned} &2\Phi\Big(-\sqrt{n} \frac{t - \mathscr{C}_1 \sum_{v \in [V]} \sum_{k=2}^{K} (a_{n,k,v} b_{n,k,v} + n^{-1/2}(a_{n,k,v} + b_{n,k,v}) + n^{-1})}{\sigma_{*,\mathrm{GSC}}} \Big) \\ &\quad + \frac{\mathscr{C}_2}{\sqrt{n}} + \sum_{v \in [V]} \sum_{k=2}^{K} (c_{n,k,v} + d_{n,k,v}) + \varepsilon, \end{aligned} \tag{20}$$

where $\sigma_{*,\mathrm{GSC}}$ is defined in (10) and Φ denotes the cumulative distribution function of the standard normal distribution.

1. Efficiency: *Under Condition ST.1, with $\hat{r}$ in Line 9 of Algorithm 1, and D_{GSC} in (8),*

$$\hat{r} - \Delta = r_* + \frac{1}{n} \sum_{i=1}^{n} D_{\mathrm{GSC}}(\boldsymbol{\ell}_*, \boldsymbol{\lambda}_*, \boldsymbol{\pi}_*, r_*)(O_i) + \mathrm{o}_p(n^{-1/2}). \tag{21}$$

For $\hat{r}$ in Line 9 of Algorithm 1†*, with D_{SC} in (9), this specializes to*

$$\hat{r} - \Delta = r_* + \frac{1}{n} \sum_{i=1}^{n} D_{\mathrm{SC}}(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*, \boldsymbol{\pi}_*, r_*)(O_i) + \mathrm{o}_p(n^{-1/2}). \tag{22}$$

2. Multiply robust consistency: *Under Condition ST.2, $\hat{r} - \Delta \xrightarrow{p} r_*$ as $n \to \infty$.*

[6] The denominator $\sum_{a \in \mathcal{S}_1'} \hat{\pi}_v^a$ is nonzero with probability tending to one exponentially.

Additionally under Condition DS.0, $\Delta = 0$ *and thus* $\hat{r}$ *is RAL and efficient under Condition* ST.1 *and is consistent for* r_* *under Condition* ST.2.

We have dropped the dependence of Δ_v and Δ on the sample size n, the nuisance estimators $(\widehat{\boldsymbol{\ell}}_v, \widehat{\boldsymbol{\lambda}}_v, \widehat{\boldsymbol{\theta}}_v, \widehat{\boldsymbol{\pi}}_v)$ and the true distribution P_* from the notation for conciseness. To illustrate the finite-sample confidence guarantee, suppose that $\|\hat{\lambda}_v^{k-1} - \lambda_*^{k-1}\|_{L^2(P_*)} = \mathrm{O}_p(a_n)$ and $\|\hat{\ell}_v^{k-1} - h_v^{k-1}\|_{L^2(P_*)} = \mathrm{O}_p(b_n)$. Such convergence rates have been established for many flexible regression or classification methods. For example, for Condition DS.0†, with d denoting the dimension of Z_{K-1} and highly adaptive lasso used to obtain $\widehat{\boldsymbol{\ell}}_v$ and $\widehat{\boldsymbol{\theta}}_v$, one has $a_n = b_n = n^{-1/4-1/\{8(d+1)\}}$ (Benkeser and van der Laan (2016)) if the nuisance functions have bounded variation norm. See Supplement S3 for more examples. Then, taking $c_n = d_n = \varepsilon/\{2V(K-1)\}$, the finite-sample guarantee becomes

$$\Pr\big(|\hat{r} - \Delta - r_*| > t\big) \le 2\Phi\biggl(-\sqrt{n}\frac{t - \mathscr{C}_1(a_n b_n + n^{-1/2}(a_n + b_n) + n^{-1})}{\sigma_{*,\mathrm{GSC}}}\biggr) + \frac{\mathscr{C}_2}{\sqrt{n}} + 2\varepsilon$$

for a different absolute constant $\mathscr{C}_1$. Thus, when the highly adaptive lasso is used, the above bound equals $2\Phi(-[\sqrt{n}t - \mathscr{C}_1 n^{-1/\{4(d+1)\}}]/\sigma_{*,\mathrm{GSC}}) + \mathscr{C}_2/\sqrt{n} + 2\varepsilon$. This leads to a nontrivial probability bound for any $t > \mathscr{C}_1 n^{-1/2-1/\{4(d+1)\}}$ such as $t \propto n^{-1/2}$. Under Conditions DS.0† and ST.1, statistical inference about r_* can be performed based on $\hat{r}$ and a consistent estimator of its influence function $D_{\mathrm{SC}}(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*, \boldsymbol{\pi}_*, r_*)$; here, a consistent estimator of the asymptotic variance of $\hat{r}$ is $\frac{1}{n}\sum_{v \in [V]} \sum_{i \in I_v} D_{\mathrm{SC}}(\widehat{\boldsymbol{\ell}}_v, \widehat{\boldsymbol{\theta}}_v, \widehat{\boldsymbol{\pi}}_v, \hat{r}_v)(O_i)^2$. Inference about r_* under the more general condition DS.0 can be conducted similarly. The results in Theorem 1 under Condition DS.0 or DS.0† can be shown using standard approaches to analyzing Z-estimators (see, e.g., Section 3.3 in van der Vaart and Wellner (1996)). However, to study the behavior of our estimator $\hat{r}$ without Condition DS.0, we need to carefully study the expansion of the mean of the estimating function D_{GSC} to identify the bias term Δ due to failure of Condition DS.0. The proof of Theorem 1 can be found in Supplement S9.3.

We also have the following immediate corollary of Theorem 1 for model comparison and model selection. Let $\ell^{(1)}$ and $\ell^{(2)}$ be two given losses. For example, for each $j \in \{1, 2\}$, $\ell^{(j)}$ may be the squared error loss $z = (x, y) \mapsto (y - f^{(j)}(x))^2$ for a given predictor $f^{(j)}$. With superscript (1) and (2) denoting quantities or functions corresponding to these two predictors, the contrast of their risk $r_*^{(1)} - r_*^{(2)}$ informs the difference between the two models' performance. It is natural to select the model with the smaller estimated risk.

COROLLARY 1. *Under Condition* DS.0, $\hat{r}^{(1)} - \hat{r}^{(2)}$ *equals*

$$r_*^{(1)} - r_*^{(2)} + \frac{1}{n}\sum_{i=1}^{n}\bigl\{D_{\mathrm{GSC}}\bigl(\boldsymbol{\ell}_*^{(1)}, \boldsymbol{\lambda}_*, \boldsymbol{\pi}_*, r_*^{(1)}\bigr)(O_i) - D_{\mathrm{GSC}}\bigl(\boldsymbol{\ell}_*^{(2)}, \boldsymbol{\lambda}_*, \boldsymbol{\pi}_*, r_*^{(2)}\bigr)(O_i)\bigr\} + \mathrm{o}_p\bigl(n^{-1/2}\bigr)$$

if Condition ST.1 *holds for both losses, and* $\hat{r}^{(1)} - \hat{r}^{(2)} \xrightarrow{p} r_*^{(1)} - r_*^{(2)}$ *if Condition* ST.2 *holds for both losses. Moreover, if* $r_*^{(1)} - r_*^{(2)} = -C/\sqrt{n}$ *for a constant* $C > 0$ *and Condition* ST.1 *holds for both losses, then the probability for the estimated sign of* $r_*^{(1)} - r_*^{(2)}$ *to be correct* $\Pr(\hat{r}^{(1)} < \hat{r}^{(2)}) \to \Phi(C/\chi)$ *(as* $n \to \infty$*) where* $\chi := (\mathbb{E}_P[\{D_{\mathrm{GSC}}(\boldsymbol{\ell}_*^{(1)}, \boldsymbol{\lambda}_*, \boldsymbol{\pi}_*, r_*^{(1)})(O) - D_{\mathrm{GSC}}(\boldsymbol{\ell}_*^{(2)}, \boldsymbol{\lambda}_*, \boldsymbol{\pi}_*, r_*^{(2)})(O)\}^2])^{1/2}$ *is the asymptotic standard deviation of the estimator* $\hat{r}^{(1)} - \hat{r}^{(2)}$.

The nuisances $\boldsymbol{\lambda}_*$ and $\boldsymbol{\pi}_*$ are invariant with respect to the loss and so need not be estimated twice for the two losses. Under Condition ST.1, inference about the risk contrast $r_*^{(1)} - r_*^{(2)}$ can be conducted based on $\hat{r}^{(1)} - \hat{r}^{(2)}$ and a consistent estimator of its asymptotic variance,

$\frac{1}{n}\sum_{v\in[V]}\sum_{i\in I_v}\{D_{\mathrm{GSC}}(\boldsymbol{\ell}_*^{(1)},\boldsymbol{\lambda}_*,\boldsymbol{\pi}_*,r_*^{(1)})(O_i)-D_{\mathrm{GSC}}(\boldsymbol{\ell}_*^{(2)},\boldsymbol{\lambda}_*,\boldsymbol{\pi}_*,r_*^{(2)})(O_i)\}^2$. In the above example of model comparison, rejection of the null hypothesis $r_*^{(1)}-r_*^{(2)}=0$ indicates a significant difference between the two predictors' performance in the target population; otherwise, the two predictors might perform similarly.

In the example of selecting the model with a smaller risk based on the estimated sign of risk difference $r_*^{(1)}-r_*^{(2)}$, consider the case where the true risk difference is small even in relatively large samples, namely the asymptotic regime where $r_*^{(1)}-r_*^{(2)}=-C/\sqrt{n}$ for a constant C as in Corollary 1. Without loss of generality, let $C>0$. If the risk difference is estimated with the nonparametric estimator, then the asymptotic probability of estimating the correct sign is asymptotically $\Phi(C/\chi_{\mathrm{np}})$ where $\chi_{\mathrm{np}}:=(\mathbb{E}_P[\{D_{\mathrm{np}}(\rho,r_*^{(1)})(O)-D_{\mathrm{np}}(\rho,r_*^{(2)})(O)\}^2])^{1/2}$ is the asymptotic standard deviation of the nonparametric estimator. Because our proposed estimator is asymptotically efficient, $\chi_{\mathrm{np}}^2\geq\chi^2$ and our proposed estimator results in a greater probability of selecting the better-performing model than using the nonparametric estimator. Similarly, consider $J+1$ risks $r_*^{(1)},\dots,r_*^{(J+1)}$ such that $r_*^{(1)}-r_*^{(j)}=-C_{j-1}/\sqrt{n}$ ($C_{j-1}>0$, $j=2,\dots,J+1$). Suppose that the asymptotic covariance matrices of $(\hat{r}^{(1)}-\hat{r}^{(2)},\dots,\hat{r}^{(1)}-\hat{r}^{(J+1)})$ and $(\hat{r}_{\mathrm{np}}^{(1)}-\hat{r}_{\mathrm{np}}^{(2)},\dots,\hat{r}_{\mathrm{np}}^{(1)}-\hat{r}_{\mathrm{np}}^{(J+1)})$ are Σ and Σ_{np}, respectively. Then the asymptotic probabilities of estimating the correct smallest risk based on these estimators are $\Pr(\mathrm{N}_J(0,\Sigma)<(C_1,\dots,C_J))$ and $\Pr(\mathrm{N}_J(0,\Sigma_{\mathrm{np}})<(C_1,\dots,C_J))$, respectively, where N_J denotes the J-dimensional normal distribution, and the events in these probabilities mean that a J-dimensional normal random vector is less than the vector $(C_1,\dots,C_J)$ entrywise; the former asymptotic probability is greater than or equal to the latter because $\Sigma\preceq\Sigma_{\mathrm{np}}$.

3.3. *Discussion on estimation of conditional mean loss function.*

In contrast to our approach in Algorithm 1†, a *direct regression method* for estimating nuisance functions $\boldsymbol{\ell}_*$ is to regress $\ell(Z)$ on covariates in the subsample with $A=0$, since $\ell_*^k(\bar{z}_k)=\mathbb{E}_Q[\ell(Z)\mid\bar{Z}_k=\bar{z}_k]$ for $\bar{z}_k\in\mathcal{Z}_k$ under Condition DS.0. Direct regression aims to estimate nuisance functions ℓ_*^k rather than h_*^k, and can also achieve efficiency under Condition DS.0† when all nuisance functions are estimated consistently at sufficient rates to satisfy Condition ST.1.

However, our sequential regression approach is advantageous in achieving multiple robustness under less stringent conditions on nuisance function estimators. We illustrate this advantage for Condition DS.0† and Algorithm 1† as an example. Note that B_k from (13) appearing in the sum (14) involves the difference between the nuisance estimator $\hat{\ell}_v^{k-1}$ and the oracle estimator h_v^{k-1}, which depends on the nuisance estimator $\hat{\ell}_v^k$ in the previous step. Each nuisance function estimator $\hat{\ell}_v^k$ (except those with indices $k=K$ and $k=2$) appears in both B_k and B_{k-1} in (14).

In the direct regression method, we might wish to achieve 2^{K-1}-robustness in the sense that the final risk estimator is consistent for r_* if, for each $k\in[K-1]$, either $\hat{\ell}_v^k$ or $\hat{\theta}_v^k$ is consistent. Consider a fixed index $j\in[K-1]$. If all nuisance estimators in $\widehat{\boldsymbol{\ell}}_v$ except $\hat{\ell}_v^j$ are consistent, then neither of $h_v^j-\hat{\ell}_v^j$ and $h_v^{j-1}-\hat{\ell}_v^{j-1}$ would converge to zero. If $\hat{\theta}_v^j$ is consistent but $\hat{\theta}_v^{j-1}$ is not, then, in general, (14) would not converge to zero, and thus the risk estimator is inconsistent. In other words, the above approach might not achieve the desired 2^{K-1}-robustness property.

In contrast, our sequential regression approach directly aims to estimate the oracle regression functions h_v^k and the estimator $\hat{\ell}_v^k$ inherits the potential bias in $\hat{\ell}_v^{k+1}$. For example, if all differences $h_v^k-\hat{\ell}_v^k$ except $h_v^j-\hat{\ell}_v^j$ converge to zero, in order to make (14) small, it is indeed sufficient for $1/(1+\hat{\theta}_v^j)$ to be consistent for $1/(1+\theta_*^j)$. It seems that such a multiple robustness property could only hold for the direct regression method under stringent or even implausible conditions on the nuisance function estimators $\hat{\ell}_v^k$.

3.4. *Consequences of the failure of Condition* DS.0. In this section, we focus on Condition DS.0† and Algorithm 1† for illustration. All results apply to the more general condition DS.0 unless otherwise stated. We first show the intriguing fact that, when Condition DS.0 fails, the conditional mean loss ℓ_*^k might not be uniquely defined even in $\mathcal{Z}_k$. The reason is that the supports of $\bar{Z}_k \mid A = a$ may be potentially mismatched across Q and $a \in \mathcal{S}_k'$ for $k \in [K]$, and moreover that for $j > k$, ℓ_*^j may not be uniquely defined.

Consider the following simple example with $K = 3$. Suppose that the support of $\bar{Z}_2 \mid A \in \mathcal{S}_3'$ is one point $\{(0,0)\}$, while the supports of $\bar{Z}_2 \mid A \in \mathcal{S}_2'$ and $\bar{Z}_2 \mid A \in \mathcal{S}_1'$ are both two points $\{(0,0),(0,1)\}$. Suppose that $\mathcal{Z}_2 = \{(0,0)\}$ and therefore $\mathcal{Z}_1 = \{0\}$. Then, ℓ_*^2 is nonuniquely defined at $(0,1)$. This nonunique definition is allowed for under Condition DS.0†. However, since $\ell_*^1(\bar{Z}_1) = \mathbb{E}_{P_*}[\ell_*^2(\bar{Z}_2) \mid \bar{Z}_1, A \in \mathcal{S}_2']$, when Condition DS.0† fails, the value of ℓ_*^1 at $0 \in \mathcal{Z}_1$ depends on the value of ℓ_*^2 at $(0,1)$, which is not uniquely defined. In other words, ℓ_*^1 is nonuniquely defined even in the support $\mathcal{Z}_1$ of $\bar{Z}_1 \mid A = 0$.

We note that this dependence of ℓ_*^1 on $\ell_*^2(\bar{z}_2)$ for $\bar{z}_2 = (0,1) \notin \mathcal{Z}_2$ is excluded under Condition DS.0†: $(0,1)$ cannot be in the support of $\bar{Z}_2 \mid A \in \mathcal{S}_2'$ by the assumption that $\bar{Z}_2 \mid Z_1 = z_1$, $A = 0$ and $\bar{Z}_2 \mid Z_1 = z_1$, $A \in \mathcal{S}_2'$ are identically distributed for $z_1 \in \mathcal{Z}_1$. This nonunique definition might also be reflected in the corresponding nuisance estimator $\widehat{\boldsymbol{\ell}}_v$: $\hat{\ell}_v^k$ might be (unintentionally) extrapolated to outside the support of $\bar{Z}_k \mid A \in \mathcal{S}_k'$ in order to obtain the estimator $\hat{\ell}_v^{k-1}$ in Line 6, Algorithm 1†. This support issue might go undetected in the estimation procedure. The oracle estimator h_v^{k-1} would also depend on how $\hat{\ell}_v^k$ is extrapolated.

Nevertheless, without Condition DS.0, our results in Section 3.2 remain valid as long as Condition ST.1 or ST.2 holds for one version of the collection of true conditional mean loss functions $\boldsymbol{\ell}_*$. For example, part 2 of Condition ST.1 would require the consistency of $\widehat{\boldsymbol{\ell}}_v$ for some version of $\boldsymbol{\ell}_*$, and $D_{\mathrm{SC}}(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*, \boldsymbol{\pi}_*, r_*)$ in (22) would depend on the particular adopted version of $\boldsymbol{\ell}_*$. The appropriate choice of $\boldsymbol{\ell}_*$ often depends on the asymptotic behavior of the nuisance estimator $\widehat{\boldsymbol{\ell}}_v$, which might heavily depend on the particular choice of the regression technique used in the sequential regression (Line 6, Algorithm 1†).

The choice of regression techniques can further affect the bias term Δ when Condition DS.0† fails. Because of the potential extrapolation when evaluating and estimating $\widehat{\boldsymbol{\ell}}_v$, the bias term Δ can have drastically different behavior for different estimators $\widehat{\boldsymbol{\ell}}_v$, even if these estimators are all consistent for some $\boldsymbol{\ell}_*$ when restricted to $\mathcal{Z}_k$. Consequently, Δ might not have a probabilistic limit, and so the estimator $\hat{r}$ can diverge.

We illustrate the behavior of Δ in the following example of concept shift in the features (DS.1), a special case of Condition DS.0†. Under the setup of this condition,

$$\Delta_v \approx \mathbb{E}_{P_*}[\mathbb{E}_{P_*}[\ell(X,Y) \mid X, A = 0]] - r_* + \mathbb{E}_{P_*}[\hat{\ell}_v^1(X)] - \mathbb{E}_{P_*}[\hat{\ell}_v^1(X) \mid A = 0]$$

where we have dropped the estimation error of order $\mathrm{O}_p(n^{-1/2})$ in estimating π_*^a with $\hat{\pi}_v^a$ in this approximation. If Condition DS.1 in fact does not hold and the difference $\mathcal{B}$ between the support of $X \mid A = 1$ and that of $X \mid A = 0$ is nonempty, the asymptotic behavior of the third term $\mathbb{E}_{P_*}[\hat{\ell}_v^1(X)]$ would depend on how the estimator $\hat{\ell}_v^1$ behaves asymptotically in $\mathcal{B}$, even if this estimator is known to be consistent for $x \mapsto \mathbb{E}_{P_*}[\ell(X,Y) \mid X = x, A = 0]$ when restricted to the support of $X \mid A = 0$. If $\hat{\ell}_v^1$ diverges in the region $\mathcal{B}$ as $n \to \infty$, our estimator $\hat{r}_v$ can diverge. This phenomenon is fundamental and cannot be resolved by, for example, using an assumption lean approach (Vansteelandt and Dukes (2022)) because it mirrors the ill-defined nuisance functions ℓ_*^k at $\bar{z}_k \notin \mathcal{Z}_k$ ($k \in [K-1]$).

In practice, mismatched supports and extrapolation of the estimator $\widehat{\boldsymbol{\ell}}_v$ caused by failure of Condition DS.0 might be detected from extreme values or even numerical errors when evaluating $\hat{\ell}_v^k$ at sample points, but such detection is not guaranteed. When target population data is observed, the above analysis of our estimator $\hat{r}$—which leverages the dataset shift condition DS.0† to gain efficiency— motivates the following result that leads to a test of whether $\hat{r}$ is consistent for r_*.

COROLLARY 2 (Testing root-n consistency of $\hat{r}$). *Under Conditions* DS.0† *and* ST.1, $\sqrt{n}(\hat{r} - \hat{r}_{\mathrm{np}})$ *is asymptotically distributed as a normal distribution with mean zero and variance* $\mathbb{E}_{P_*}[\{D_{\mathrm{SC}}(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*, \boldsymbol{\pi}_*, r_*)(O) - D_{\mathrm{np}}(\Pi_*, r_*)(O)\}^2] = \sigma^2_{*,\mathrm{np}} - \sigma^2_{*,\mathrm{SC}}$, *as* $n \to \infty$.

This corollary is implied by Theorem 1 and the orthogonality between $D_{\mathrm{SC}}(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*, \boldsymbol{\pi}_*, r_*)$ -$D_{\mathrm{np}}(\Pi_*, r_*)$ and $D_{\mathrm{SC}}(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*, \boldsymbol{\pi}_*, r_*)$ under Condition DS.0†. This result might not hold for the more general condition DS.0 due to potential lack of a nonparametric estimator $\hat{r}_{\mathrm{np}}$. A specification test (Hausman (1978)) of the dataset shift condition DS.0† can be constructed based on the two estimators $\hat{r}_{\mathrm{np}}$ and $\hat{r}$ along with their respective standard errors SE_1 and SE_2. Under Condition ST.1 and the null hypothesis Condition DS.0†, the test statistic[7] $(\hat{r} - \hat{r}_{\mathrm{np}})/\{(\mathrm{SE}_1^2 - \mathrm{SE}_2^2)^{1/2}\}$ is approximately distributed as N(0, 1) in large samples; in contrast, if Condition DS.0† does not hold, $\hat{r}$ is generally inconsistent for r_* and thus the test statistic diverges as $n \to \infty$.

The aforementioned test of Condition DS.0† may be underpowered because only one loss function ℓ is considered. For example, it is possible to construct a scenario where Condition DS.0† fails, while the loss function ℓ and the nuisance function estimators $\widehat{\boldsymbol{\ell}}_v$ are chosen such that $\Delta = \mathrm{o}_p(n^{-1/2})$. In this case, the asymptotic power of the aforementioned test is no greater than the asymptotic type I error rate. A somewhat contrived construction is to set $\widehat{\boldsymbol{\ell}}_v$ as one version of the true conditional mean loss $\boldsymbol{\ell}_*$ and choose P_* such that $\mathbb{E}_{P_*}[\ell_*^k(\bar{Z}_k) \mid \bar{Z}_{k-1}, A = 0] = \ell_*^{k-1}(\bar{Z}_{k-1})$ holds. This implies that $\Delta = 0$, while Condition DS.0† can fail, for example, due to the heteroskedasticity of the residuals $\ell_*^k(\bar{Z}_k) - \ell_*^{k-1}(\bar{Z}_{k-1})$. Such phenomena have also been found in specification tests for generalized method of moments (Newey (1985)). More powerful tests of conditional independence that do not suffer from the above phenomenon have been proposed in other settings (see, e.g., Doran et al. (2014), Hu and Lei (2023), Shah and Peters (2020), Zhang et al. (2011), etc.).

However, since $\hat{r}$ might still be root-n consistent for r_* even if Condition DS.0† does not hold, the above test should be interpreted as a test of the null hypothesis that $\hat{r}$ is root-n consistent for r_*, a weaker null hypothesis than conditional independence (Condition DS.0†). Nevertheless, this weaker null hypothesis is meaningful when the risk r_* is the estimand of interest.

REMARK 4. It is possible to adopt our proposed estimator when the relevant source populations $\mathcal{S}_k$ in Condition DS.0 or DS.0† are not known *a priori* but can be selected based on data. In this case, the user can split the data into two folds, use fold 1 to select $\mathcal{S}_k$, and finally compute our estimator on fold 2. We leave a more thorough study of such approaches to future work.

For the following Sections 4 and 5, we focus on risk estimation under one of the four popular dataset shift conditions DS.1–DS.4. The special structures of these conditions will be further exploited in the estimation procedure, leading to additional simplifications, more flexibility in estimation, and potentially more robustness. For example, for Conditions DS.1 and DS.2, $K = 2$ and $\mathcal{S}_2$ is known to be empty, and we will show that estimators with better robustness properties than part 2 of Theorem 1 can be constructed; for Conditions DS.3

[7] It is viable to use a variant statistic with the denominator replaced by an asymptotic variance estimator based on the influence function $D_{\mathrm{SC}}(\boldsymbol{\ell}_*, \boldsymbol{\theta}_*, \boldsymbol{\pi}_*, r_*) - D_{\mathrm{np}}(\Pi_*, r_*)$.

and DS.4, $K = 2$ and $\mathcal{S}_1$ is known to be empty, and the estimation procedure described in Algorithm 1† can be simplified.

Conditions DS.1 and DS.2 are identical up to switching the roles of X and Y; the same holds for the other two conditions DS.3 and DS.4. Therefore, we study Conditions DS.1 and DS.3 in the main body and present results for Conditions DS.2 and DS.4 in Supplement S2.

3.5. *More general dataset shift condition.* Li, Gilbert and Luedtke (2023) considered a condition more general than DS.0. In addition to Condition DS.0, for each $k \in [K]$, their setting also allows for a population index subset $\mathcal{V}_k \subset \mathcal{A}$ such that, for each $a \in \mathcal{V}_k$, and all z_k, $\bar{z}_{k-1}$,

$$\mathrm{d}P_{Z_k|\bar{Z}_{k-1},A=a}/\mathrm{d}Q_{Z_k|\bar{Z}_{k-1}}(z_k \mid \bar{z}_{k-1}) \propto w_{k,a}(\bar{z}_k; \beta_{k,a})$$

for some *unknown* finite-dimensional parameter $\beta_{k,a}$ and a known tilting function $w_{k,a}$. This allows for dataset shift up to unknown finite-dimensional parameters. For instance, this includes the following cases:

1. *Example 1. Covariate shift up to exponential tilting.* We can consider a generalization of Condition DS.3, where the density of $Y \mid X = x$, $A = 1$ is proportional to the density of $Y \mid X = x$, $A = 0$ multiplied by $w(x, y; \beta) = \exp([x, y]^\top \beta)$, for all x, y.
2. *Example 2. Covariate shift with truncation.* Another generalization of Condition DS.3 allows Y in the source data to be truncated (Jewell (1985), Bickel et al. (1993), Bhattacharya, Chernoff and Yang (2007)); for example, Y is only observed when it is above an unknown threshold β. In this case, the density of $Y \mid X = x$, $A = 1$ is proportional to the density of $Y \mid X = x$, $A = 0$ multiplied by $w(x, y; \beta) = \mathbb{1}(y \geq \beta)$, for all x, y.
3. *Example 3. Covariate shift with clipping.* Suppose that Y is integer-valued, such as a count variable. Condition DS.3 can be generalized to allow Y in the source data to be clipped at a threshold B; that is, if the true outcome is above B, the observed Y equals B. In this case, the density (with respect to the counting measure) of $Y \mid X = x$, $A = 1$ is proportional to the density of $Y \mid X = x$, $A = 0$ multiplied by $w(x, y; \beta) = \mathbb{1}(y < B) + \mathbb{1}(y = B)\exp(\beta)$ for an unknown normalizing parameter β, for all x, y.

Concept shift and label shift can be extended similarly; and these constructions also clearly apply to more the general sequential conditionals setting. For this more general condition with unknown finite-dimensional parameters, Li, Gilbert and Luedtke (2023) derived influence functions that have smaller variances those not using the source data from $\mathcal{V}_k$.

Let D^* be such an influence function with a reduced variance. Similarly to the cross-fitting strategy in Algorithm 1†, for each fold $v \in [V]$, let $\hat{D}_v$ denote an estimator of D_* based on data out of fold v, which involves an estimator $\hat{\ell}_v^1$ of ℓ_*^1. With data split into V folds and $P^{n,v}$ denoting the empirical distribution of data in fold v ($v \in [V]$), consider a cross-fit one-step estimator

$$n^{-1} \sum_{v \in [V]} |I_v| \{ P^{n,v}_{Z_1|A \in \mathcal{S}_1'} \hat{\ell}_v^1 + P^{n,v} \hat{D}_v \}.$$

If all nuisance functions are estimated well in the sense that $\|\hat{D}_v - D_*\|_{L^2(P_0)} = \mathrm{o}_p(1)$ and $\mathscr{R}_v := P^{n,v}_{Z_1|A \in \mathcal{S}_1'} \hat{\ell}_v^1 - r_* + P_* \hat{D}_v = \mathrm{o}_p(n^{-1/2})$ for all $v \in [V]$, then the one-step estimator $P^{n,v}_{Z_1|A \in \mathcal{S}_1'} \hat{\ell}_v^1 + P^{n,v} \hat{D}_v$ for fold v equals

$$r_* + P^{n,v} D_* + (P^{n,v} - P_*)(\hat{D}_v - D_*) + \mathscr{R}_v = r_* + P^{n,v} D_* + \mathrm{o}_p(n^{-1/2}).$$

With P^n denoting the empirical distribution of the entire data, the cross-fit one-step estimator equals $r_* + P^n D_* + \mathrm{o}_p(n^{-1/2})$, that is, it is asymptotically normal with a *reduced asymptotic variance compared to the nonparametric estimator*. We leave studying multiple robustness for future work.

4. Concept shift in the features.

4.1. *Efficiency bound.* We first present the efficient influence function for the risk r_* under concept shift in the features, where $X \perp\!\!\!\perp A$ (DS.1). To do so, define $\mathcal{E}_* : x \mapsto \mathbb{E}_{P_*}[\ell(X, Y) \mid X = x, A = 0]$, the conditional risk function in the target population. Recall that ρ_* denotes $P_*(A = 0)$. For scalars $\rho \in (0, 1)$, $r \in \mathbb{R}$, and a function $\mathcal{E} : \mathcal{X} \to \mathbb{R}$, define

$$D_{\text{Xcon}}(\rho, \mathcal{E}, r) : o = (x, y, a) \mapsto \frac{1-a}{\rho}\{\ell(x, y) - \mathcal{E}(x)\} + \mathcal{E}(x) - r. \tag{23}$$

We next present the efficient influence function, which is implied by the efficiency bound under Condition DS.0† from (9), along with the efficiency gain of an efficient estimator.

COROLLARY 3. *Under Condition* DS.1, *the efficient influence function for the risk r_* from* (1) *is $D_{\text{Xcon}}(\rho_*, \mathcal{E}_*, r_*)$ with D_{Xcon} from* (23). *Thus, the smallest possible normalized limiting variance of a sequence of RAL estimators is $\sigma^2_{*,\text{Xcon}} := \mathbb{E}_{P_*}[D_{\text{Xcon}}(\rho_*, \mathcal{E}_*, r_*)(O)^2]$.*

COROLLARY 4. *Under conditions of Corollary* 3, *the relative efficiency gain from using an efficient estimator is*

$$1 - \frac{\sigma^2_{*,\text{Xcon}}}{\sigma^2_{*,\text{np}}} = \frac{(1-\rho_*)\mathbb{E}_{P_*}[(\mathcal{E}_*(X) - r_*)^2]}{\mathbb{E}_{P_*}[\mathbb{E}_{P_*}[\{\ell(X,Y) - \mathcal{E}_*(X)\}^2 \mid A = 0, X]] + \mathbb{E}_{P_*}[\{\mathcal{E}_*(X) - r_*\}^2]}.$$

Since $\mathcal{E}_*(X) = \mathbb{E}_{P_*}[\ell(X, Y) \mid X, A = 0]$, conditioning on $A = 0$ throughout, recall the tower rule decomposition of the variance of loss $\ell(X, Y)$:

$$\mathbb{E}_{P_*}[\{\ell(X,Y) - r_*\}^2] = \underbrace{\mathbb{E}_{P_*}[\mathbb{E}_{P_*}[\{\ell(X,Y) - \mathcal{E}_*(X)\}^2 \mid X]]}_{\text{variability not just due to } X} + \underbrace{\mathbb{E}_{P_*}[\{\mathcal{E}_*(X) - r_*\}^2]}_{\text{variability due to } X \text{ alone}}.$$

By Corollary 4, more relative efficiency gain is achieved for estimating the true risk r_* at a data-generating distribution P_* with the following properties:

1. The proportion ρ_* of target population data is small.
2. In the target population, the proportion of variance of $\ell(X, Y)$ due to X alone is large compared to that not just due to X but rather also due to Y.

REMARK 5. We illustrate the second property in an example with squared error loss $\ell(x, y) = (y - f(x))^2$ for a given predictor f. We consider the target population and condition on $A = 0$ throughout. Let $\mu_* : x \mapsto \mathbb{E}_{P_*}[Y \mid X = x]$ be the oracle predictor and suppose that $Y = \mu_*(X) + \varepsilon$ for independent noise $\varepsilon \perp\!\!\!\perp X$. In this case, the variance of $\ell(X, Y)$ not due to X is determined by the random noise ε, while that due to X is determined by the bias $f - \mu_*$. Therefore, the proportion of variance of $\ell(X, Y)$ not due to X would be large if the given predictor f is far from the oracle predictor μ_* heterogeneously. In a related paper, Azriel et al. (2021) showed that, for linear regression under semisupervised learning, namely concept shift in the features, there is efficiency gain only if the linear model is misspecified. Our observation is an extension to more general risk estimation problems.

4.2. *Cross-fit risk estimator.* In this section, we present our proposed estimator of the risk r_*, along with its theoretical properties. This estimator is described in Algorithm 2. This algorithm is the special case of Algorithm 1† with simplifications implied by Condition DS.1.

We next present the efficiency of the estimator $\hat{r}_{\text{Xcon}}$, along with its fully robust asymptotic linearity: $\hat{r}_{\text{Xcon}}$ is asymptotically linear even if the nuisance function $\mathcal{E}_*$ is estimated inconsistently. This robustness property is stronger and more desirable than that stated in part 2 of

Algorithm 2 Cross-fit estimator of risk r_* under Condition DS.1, concept shift in the features

```
Require: Data $\{O_i = (Z_i, A_i)\}_{i=1}^n$, number $V$ of folds, regression estimation method for $\mathcal{E}_*$
1: Randomly split all data (from both populations) into $V$ folds. Let $I_v$ be the indices of data points in fold $v$.
2: for $v \in [V]$ do Estimate $\mathcal{E}_*$ by $\hat{\mathcal{E}}^{-v}$ using data out of fold $v$.
3: for $v \in [V]$ do          ▷ (Obtain an estimating-equation-based estimator for fold $v$)
4:     With $\hat{\rho}^v := |I_v|^{-1} \sum_{i \in I_v} \mathbb{1}(A_i = 0)$, set
   (24)   $\hat{r}^v_{\mathrm{Xcon}} := \frac{1}{|I_v|} \sum_{i \in I_v} \left\{ \frac{\mathbb{1}(A_i = 0)}{\hat{\rho}^v} \left[\ell(X_i, Y_i) - \hat{\mathcal{E}}^{-v}(X_i)\right] + \hat{\mathcal{E}}^{-v}(X_i) \right\}.$
5: Obtain the cross-fit estimator: $\hat{r}_{\mathrm{Xcon}} := \frac{1}{n} \sum_{v=1}^V |I_v| \hat{r}^v_{\mathrm{Xcon}}$.
```

Theorem 1, a multiply robust consistency. Moreover, the efficiency of $\hat{r}_{\mathrm{Xcon}}$ only relies on the consistency of the nuisance estimator $\hat{\mathcal{E}}^{-v}$ with no requirement on its convergence rate. This condition is also weaker than Condition ST.1, which is required by part 1 of Theorem 1. The proof of this result can be found in Supplement S9.4.

THEOREM 2 (Efficiency and fully robust asymptotic linearity of $\hat{r}_{\mathrm{Xcon}}$). *Suppose that there exists a function $\mathcal{E}_\infty \in L^2(P_*)$ such that* $\max_{v \in [V]} \|\hat{\mathcal{E}}^{-v} - \mathcal{E}_\infty\|_{L^2(P_*)} = \mathrm{o}_p(1)$. *Under Condition* DS.1, *the sequence of estimators $\hat{r}_{\mathrm{Xcon}}$ in Line* 5 *of Algorithm* 2 *is RAL: with r_* from* (1) *and D_{Xcon} from* (23), *$\hat{r}_{\mathrm{Xcon}}$ equals*

$$r_* + \frac{1}{n} \sum_{i=1}^n \left\{ D_{\mathrm{Xcon}}(\rho_*, \mathcal{E}_\infty, r_*)(O_i) + \frac{\mathbb{E}_{P_*}[\mathcal{E}_\infty(X)] - r_*}{\rho_*} (1 - A_i - \rho_*) \right\} + \mathcal{B}, \tag{25}$$

where

$$\mathcal{B} := \sum_{v \in [V]} \frac{|I_v|}{n} \left\{ \frac{\hat{\rho}^v - \rho_*}{\hat{\rho}^v} P_*(\hat{\mathcal{E}}^{-v} - \mathcal{E}_\infty) \right.$$
$$\left. + (P^{n,v} - P_*)\{D_{\mathrm{Xcon}}(\hat{\rho}^{-v}, \hat{\mathcal{E}}^{-v}, \hat{r}^v_{\mathrm{Xcon}}) - D_{\mathrm{Xcon}}(\rho_*, \mathcal{E}_\infty, r_*)\} \right\} = \mathrm{o}_p(n^{-1/2}).$$

Moreover, if $\mathcal{E}_*$ is estimated consistently, namely $\mathcal{E}_\infty = \mathcal{E}_*$, then $\hat{r}_{\mathrm{Xcon}}$ is efficient:

$$\hat{r}_{\mathrm{Xcon}} = r_* + \frac{1}{n} \sum_{i=1}^n D_{\mathrm{Xcon}}(\rho_*, \mathcal{E}_*, r_*)(O_i) + \mathrm{o}_p(n^{-1/2}). \tag{26}$$

REMARK 6 (Estimation of ρ_*). It is possible to replace the in-fold estimator $\hat{\rho}^v$ of ρ_* with an out-of-fold estimator in Algorithm 2. Unlike Algorithm 1[†], this would lead to a different influence function when the nuisance function $\mathcal{E}_*$ is estimated inconsistently in Theorem 2. In this case, the influence function of $\hat{r}_{\mathrm{Xcon}}$ from Theorem 2 cannot be used to construct asymptotically valid confidence intervals if $\mathcal{E}_*$ is estimated inconsistently.

REMARK 7 (A semiparametric perspective on prediction-powered inference). Our proposed estimator is distantly related to the work of Angelopoulos et al. (2023). Angelopoulos et al. (2023) studied the estimation of and inferences about a risk minimizer with the aid of an arbitrary predictor under concept shift (Condition DS.1). Their proposed risk estimator is essentially a special variant of Algorithm 2 without cross-fitting and with a fixed given estimator of $\mathcal{E}_*$. Theorem 2 provides another perspective on why their proposed method is valid for an *arbitrary* given nuisance estimator of the true conditional mean risk $\mathcal{E}_*$ and improves efficiency when the given estimator is close to the truth $\mathcal{E}_*$.

4.3. *Simulation.* We illustrate Theorem 2 and Corollary 4 in a simulation study. We consider the application of estimating the mean squared error (MSE) of a given predictor f that predicts the outcome Y given input covariate X; that is, we take $\ell(x, y) = (y - f(x))^2$. We consider five scenarios:

(A) The predictor f is identical to the oracle predictor and the additive noise is homoskedastic. According to Corollary 4 and Remark 5, there should be no efficiency gain from using our proposed estimator $\hat{r}_{\text{Xcon}}$ compared to the nonparametric estimator $\hat{r}_{\text{np}}$.

(B) The predictor f is a good linear approximation to the oracle predictor. This scenario may occur if the given predictor is fairly close to the truth.

(C) The predictor f substantially differs from the oracle predictor. This scenario may occur if the given predictor has poor predictive power, possibly because of inaccurate tuning or using inappropriate domain knowledge in the training process.

(D) The predictor f substantially differs from the oracle predictor and the outcome Y is deterministic given X. According to Corollary 4, there is a large efficiency gain from using our proposed estimator $\hat{r}_{\text{Xcon}}$ compared to the nonparametric estimator $\hat{r}_{\text{np}}$.

(E) Condition DS.1 does not hold.

Scenarios A and D are extreme cases designed for sanity checks, while Scenarios B and C are intermediate and more realistic. Scenario E is a relatively realistic case where the assumed dataset shift condition fails and is designed to check the robustness against assuming the wrong dataset shift condition.

More specifically, the data is generated as follows. We first generate the covariate $X = (X_1, X_2, X_3)$ from a trivariate normal distribution with mean zero and identity covariance matrix. For Scenarios A–D, where Condition DS.1 holds, we generate the population indicator A from Bernoulli(0.9) independent of X. In other words, $\rho_* = 10\%$ of data points are from the target population and the other 90% of data points are from the source population. The label in the source population, namely with $A = 1$, is treated as missing as it is not assumed to contain any information about the target population. The label Y in the target population, namely with $A = 0$, is generated depending on the scenario as follows: (A) $Y \mid X = x,\ A = 0 \sim \mathrm{N}(\mu_*(x), 5^2)$; (B) & (C): $Y \mid X = x,\ A = 0 \sim \mathrm{N}(\mu_*(x), 1)$; (D): $Y = \mu_*(X)$, where

$$\mu_*(x) = x_1 + x_2 + x_3 + 0.4x_1x_3 - 0.5x_2x_3 + \sin(x_1 + x_3). \tag{27}$$

We set different predictors f for these scenarios:

(A) f is the truth μ_*;

(B) f is a linear function close to the best linear approximation to μ_* in $L^2(P_*)$-sense: $f(x) = 1.4x_1 + x_2 + 1.4x_3$;

(C) f substantially differs from μ_*: $f(x) = -1 - 3x_1 + 0.5x_3$;

(D) f substantially differs from μ_*: $f(x) = x_1$.

For Scenario E, where Condition DS.1 does not hold, we include dependence of A on X by generating A as

$$A \mid X = x \sim \mathrm{Bernoulli}\big(\mathrm{expit}\big\{\cos(x_1 + x_2x_3) + 2x_1^2x_2^2 + 3|x_1x_3| + |x_2|(0.5 - x_3)\big\}\big).$$

The resulting proportion ρ_* of target population data is around 10%, similar to the other scenarios. The outcome Y is generated in the same way as Scenarios B and C. We set the fixed predictor f to be the same as in Scenario B.

We consider the following three estimators: `np`: the nonparametric estimator $\hat{r}_{\text{np}}$ in (2); `Xconshift`: our estimator $\hat{r}_{\text{Xcon}}$ from Line 5 of Algorithm 2 with a consistent estimator of $\mathcal{E}_*$; `Xconshift,mis.E`: $\hat{r}_{\text{Xcon}}$ with an inconsistent estimator of $\mathcal{E}_*$. To estimate nuisance

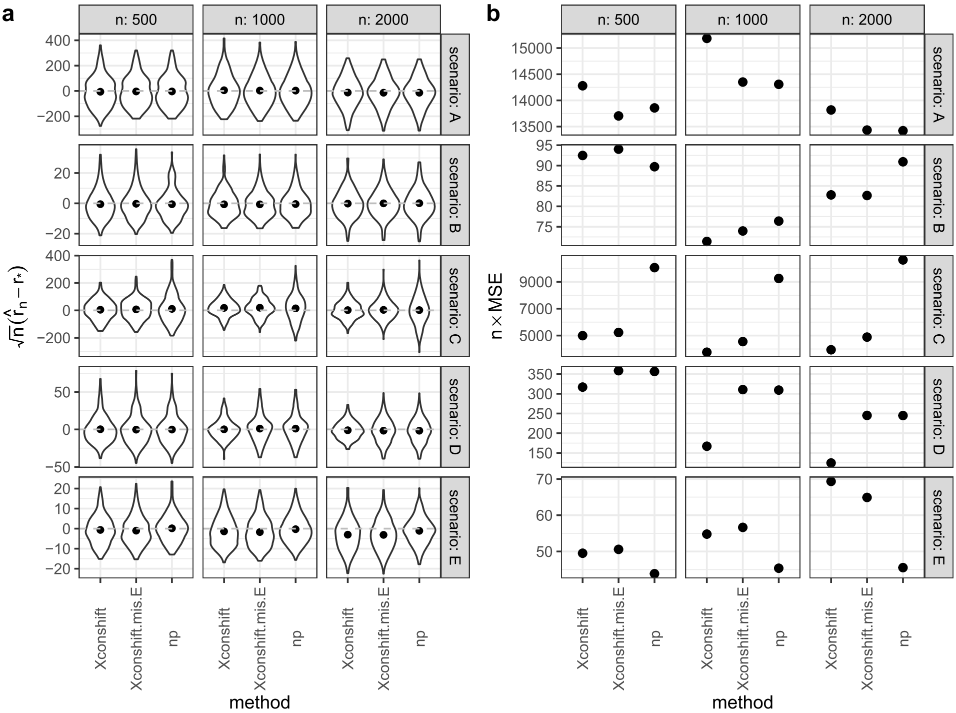

FIG. 2. (*a*) *Sampling distribution of the scaled difference between MSE estimators and the true MSE in the four scenarios under concept shift in the features. The point stands for the empirical average in Monte Carlo simulations.* (*b*) *Monte Carlo estimate of the scaled mean squared error of the estimators.*

functions consistently, we use Super Learner (van der Laan, Polley and Hubbard (2007)) whose library consists of generalized linear model, generalized additive model (Hastie and Tibshirani (1990)), generalized linear model with lasso penalty (Hastie, Buja and Tibshirani (1995), Tibshirani (1996)), and gradient boosting (Mason et al. (1999), Friedman (2001, 2002), Chen and Guestrin (2016)) with various combinations of tuning parameters. This library contains highly flexible machine learning methods and is likely to yield consistent estimators of the nuisance function. Super Learner is an ensemble learner that performs almost as well as the best learner in the library. To estimate nuisance functions inconsistently, we take the estimator as a fixed function that differs from the truth for the two extreme scenarios (A and D) for a sanity check, and drop gradient boosting from the above library for the other two relatively realistic scenarios (B and C). Since neither of generalized linear model (with or without lasso penalty) and generalized additive model is capable of capturing interactions, dropping gradient boosting would yield inconsistent estimators of the nuisance function $\mathcal{E}_*$. We consider sample sizes $n \in \{500, 1000, 2000\}$ and run 200 Monte Carlo experiments for each combination of the sample size and the scenario.

Figure 2 presents the sampling distribution of the scaled difference between the three estimators of the MSE and the true MSE. When Condition DS.1 holds, all three estimators appear close to normal and centered around the truth, demonstrating Theorem 2. The variance of `Xconshift` is much smaller than that of `np` in both Scenarios C and D for sample sizes 1000 and 2000, indicating a large efficiency gain; in the other two scenarios, the variance of these two estimators is comparable. These results are consistent with Corollary 4 and Remark 5. When Condition DS.1 does not hold (Scenario E), our proposed estimator appears consistent, indicating that $\hat{r}_{\text{Xcon}}$ might be robust against moderate failures of the concept shift condition DS.1.

5. Full-data covariate shift.

5.1. *Efficiency bound.* Let $g_* : x \mapsto P_*(A = 0 \mid X = x)$ be the propensity score function for the target population and $\mathcal{L}_* : x \mapsto \mathbb{E}_{P_*}[\ell(X, Y) \mid X = x]$ be the conditional risk function. Under Condition DS.3, $\mathcal{L}_*(x) = \mathbb{E}_{P_*}[\ell(X, Y) \mid X = x, A = 0] = \mathbb{E}_{P_*}[\ell(X, Y) \mid X = x, A = 1]$. For scalars $\rho \in (0, 1)$, $r \in \mathbb{R}$ and functions $g, \mathcal{L} : \mathcal{X} \to \mathbb{R}$, define

$$D_{\mathrm{cov}}(\rho, g, \mathcal{L}, r) : o = (x, y, a) \mapsto \frac{g(x)}{\rho}\{\ell(x, y) - \mathcal{L}(x)\} + \frac{1 - a}{\rho}\{\mathcal{L}(x) - r\}. \tag{28}$$

We have the following efficient influence function—implied by the efficiency bound under Condition DS.0† from (9)—as well as the associated relative efficiency gain.

COROLLARY 5. *Under Condition DS.3, if g_* is bounded away from zero for almost every x in the support of $X \mid A = 0$, the efficient influence function for the risk r_* is $D_{\mathrm{cov}}(\rho_*, g_*, \mathcal{L}_*, r_*)$. Thus, the smallest possible limiting normalized asymptotic variance for a sequence of RAL estimators is $\sigma^2_{*,\mathrm{cov}} := \mathbb{E}_{P_*}[D_{\mathrm{cov}}\ (\rho_*, g_*, \mathcal{L}_*, r_*)(O)^2]$.*

COROLLARY 6. *Under conditions of Corollary 5, the relative efficiency gain from using an efficient estimator is*

$$1 - \frac{\sigma^2_{*,\mathrm{cov}}}{\sigma^2_{*,\mathrm{np}}} = \frac{\mathbb{E}[g_*(X)(1 - g_*(X))\mathbb{E}_{P_*}[\{\ell(X, Y) - \mathcal{L}_*(X)\}^2 \mid X]]}{\mathbb{E}_{P_*}[g_*(X)\mathbb{E}_{P_*}[\{\ell(X, Y) - \mathcal{L}_*(X)\}^2 \mid X]] + \mathbb{E}_{P_*}[g_*(X)\{\mathcal{L}_*(X) - r_*\}^2]}.$$

By Corollary 6, more efficiency gain is achieved for P_* when the following hold:

1. The propensity score g_* is close to zero. If the covariate distribution in the source population covers that in the target population, then the proportion of source data is large.
2. The proportion of variance of $\ell(X, Y)$ not due to X is large compared to that due to X.

The second property is the opposite of the implication of Corollary 4 under concept shift in the features. Therefore, in the illustrating example in Remark 5, the efficiency gain under covariate shift would be large if the given predictor f is close to the oracle predictor μ_*.

5.2. *Cross-fit risk estimator.* We propose to use a cross-fit estimator based on estimating equations, as described in Algorithm 3. This algorithm is the special case of Algorithm 1† with simplifications implied by Condition DS.3.

Algorithm 3 Cross-fit estimator of risk r_* under full-data covariate shift condition DS.3

```
Require: Data $\{O_i = (Z_i, A_i)\}_{i=1}^n$, number $V$ of folds, classifier to estimate $g_*$, regression
    estimation method for $\mathcal{L}_*$
Randomly split all data (from both populations) into $V$ folds. Let $I_v$ be the indices of for
    fold $v$.
for $v \in [V]$ do Estimate $(g_*, \mathcal{L}_*)$ by $(\hat{g}^{-v}, \hat{\mathcal{L}}^{-v})$ using data out of fold $v$.
for $v \in [V]$ do                    ▷ (Obtain an estimating-equation-based estimator for fold $v$)
    With $\hat{\rho}^v := \frac{1}{|I_v|}\sum_{i \in I_v} \mathbb{1}(A_i = 0)$, set
(29)    $\hat{r}^v_{\mathrm{cov}} := \frac{1}{\hat{\rho}^v |I_v|} \sum_{i \in I_v} \{\hat{g}^{-v}(X_i)[\ell(X_i, Y_i) - \hat{\mathcal{L}}^{-v}(X_i)] + \mathbb{1}(A_i = 0)\hat{\mathcal{L}}^{-v}(X_i)\}.$
Obtain the cross-fit estimator: $\hat{r}_{\mathrm{cov}} := \frac{1}{n}\sum_{v=1}^V |I_v| \hat{r}^v_{\mathrm{cov}}$.
```

Compared to our proposed estimator $\hat{r}_{\text{Xcon}}$ for concept shift, the estimator $\hat{r}_{\text{cov}}$ involves two nuisance functions rather than one. As we will show next, in contrast to $\hat{r}_{\text{Xcon}}$, the efficiency of $\hat{r}_{\text{cov}}$ is based on sufficiently fast convergence rates—rather than consistency alone—of the nuisance function estimators, similarly to Theorem 1.

CONDITION ST.3 (Sufficient rate of convergence for nuisance estimators). It holds that

$$\max_{v\in[V]}\left|\int(\hat{g}^{-v}-g_*)(\hat{\mathcal{L}}^{-v}-\mathcal{L}_*)\,\mathrm{d}P_*\right| = \mathrm{o}_p(n^{-1/2}),$$
$$\max_{v\in[V]}\|\hat{g}^{-v}-g_*\|_{L^2(P_*)} = \mathrm{o}_p(1), \qquad \max_{v\in[V]}\|\hat{\mathcal{L}}^{-v}-\mathcal{L}_*\|_{L^2(P_*)} = \mathrm{o}_p(1).$$

This condition implies Condition ST.1 under Condition DS.3; but is perhaps a bit more interpretable. This leads to the efficiency of $\hat{r}_{\text{cov}}$, whose proof has the same spirit as part 1 of Theorem 1.

COROLLARY 7 (Efficiency of $\hat{r}_{\text{cov}}$). *Under Condition DS.3, with r_* from (1) and D_{cov} from (28), the estimator $\hat{r}_{\text{cov}}$ in Line 5 of Algorithm 3 has the following finite-sample expansion*:

$$\begin{aligned}\hat{r}_{\text{cov}} - r_* = &\sum_{v\in[V]}\frac{|I_v|\rho_*}{n\hat{\rho}^v}(P^{n,v}-P_*)D_{\text{cov}}(\rho_*,g_*,\mathcal{L}_*,r_*)\\ &+\sum_{v\in[V]}\frac{|I_v|}{n\hat{\rho}^v}(P^{n,v}-P_*)\{\hat{g}^{-v}(\ell-\hat{\mathcal{L}}^{-v})-g_*(\ell-\mathcal{L}_*)\}\\ &-\sum_{v\in[V]}\frac{|I_v|}{n}P_*(\hat{g}^{-v}-g_*)(\hat{\mathcal{L}}^{-v}-\mathcal{L}_*).\end{aligned}$$

Additionally, under Condition ST.3, $\hat{r}_{\text{cov}}$ is regular and efficient:

$$\hat{r}_{\text{cov}} = r_* + \frac{1}{n}\sum_{i=1}^{n}D_{\text{cov}}(\rho_*,g_*,\mathcal{L}_*,r_*)(O_i) + \mathrm{o}_p(n^{-1/2}).$$

Therefore, with $\sigma^2_{*,\text{cov}}$ from (5), $\sqrt{n}(\hat{r}_{\text{cov}}-r_*)\xrightarrow{d}\mathrm{N}(0,\sigma^2_{*,\text{cov}})$.

Another difference between $\hat{r}_{\text{cov}}$ and $\hat{r}_{\text{Xcon}}$ is that $\hat{r}_{\text{cov}}$ is doubly robust consistent, as implied by part 2 of Theorem 1, but not fully robust asymptotically linear. The doubly robust consistency of $\hat{r}_{\text{cov}}$ relies on the following condition that corresponds to Condition ST.2 for sequential conditionals and is weaker than Condition ST.3. The comparison between these conditions is similar to that between Conditions ST.1 and ST.2.

CONDITION ST.4 (Consistent estimation of one nuisance function). It holds that

$$\max_{v\in[V]}\left|\int(\hat{g}^{-v}-g_*)(\hat{\mathcal{L}}^{-v}-\mathcal{L}_*)\,\mathrm{d}P_*\right| = \mathrm{o}_p(1),$$
$$\max_{v\in[V]}\|\hat{g}^{-v}-g_*\|_{L^2(P_*)} = \mathrm{O}_p(1), \qquad \max_{v\in[V]}\|\hat{\mathcal{L}}^{-v}-\mathcal{L}_*\|_{L^2(P_*)} = \mathrm{O}_p(1).$$

We obtain the doubly robust consistency for $\hat{r}_{\text{cov}}$, a corollary of part 2 of Theorem 1.

COROLLARY 8 (Double robustness of $\hat{r}_{\text{cov}}$). *Under Conditions DS.3 and ST.3, with r_* from (1), the sequence of estimators $\hat{r}_{\text{cov}}$ from Line 5 of Algorithm 3 is consistent for r_*.*

The differences between the asymptotic properties of $\hat{r}_{\mathrm{Xcon}}$ and $\hat{r}_{\mathrm{cov}}$ are due to the differences between the dataset shift conditions. Covariate shift DS.3 is independence conditional on covariates, while concept shift DS.1 is marginal independence. One aspect in which they differ is that conditional independence is often more difficult to test than marginal independence (Shah and Peters (2020)).

When covariate shift DS.3 holds, compared to the nonparametric estimator $\hat{r}_{\mathrm{np}}$, our proposed estimator $\hat{r}_{\mathrm{cov}}$ has advantages and limitations. In terms of efficiency, $\hat{r}_{\mathrm{cov}}$ may achieve efficiency gains when both nuisance functions are estimated consistently. In terms of robustness, $\hat{r}_{\mathrm{np}}$ does not require estimating any nuisance function and is therefore fully robust asymptotically linear; in contrast, $\hat{r}_{\mathrm{cov}}$ is only doubly robust consistent but not fully robust. A natural question is whether there exists a regular estimator that is fully robust asymptotically linear and also attains the efficiency bound under reasonable conditions, similarly to $\hat{r}_{\mathrm{Xcon}}$ under concept shift. Unfortunately, as the following result shows, such estimators do not exist under the common parametrizations $(P_X, P_{A|X}, P_{Y|X})$ and $(P_A, P_{X|A}, P_{Y|X})$ of the distribution P.

LEMMA 1. *Suppose that the covariate shift condition* DS.3 *holds, but no further assumptions on P_* are made. Under the parametrization $(P_X, P_{A|X}, P_{Y|X})$ of a distribution P, suppose that for all P_*,* $\mathrm{IF}(P_{*,X}, P_{A|X}, P_{Y|X}, r_*)$ *is an influence function for estimating r_* at P_*, for arbitrary $(P_{A|X}, P_{Y|X})$. Then, we have that*

$$\mathrm{IF}(P_{*,X}, P_{A|X}, P_{Y|X}, r_*) = D_{\mathrm{np}}(\rho_*, r_*).$$

A similar result holds under the parametrization $(P_A, P_{X|A}, P_{Y|X})$ of P.

Therefore, if a regular estimator of r_* is fully robust asymptotically linear under either of the above parametrizations, that is, its influence function satisfies the assumptions for IF in either case of Lemma 1, then its influence function must be $D_{\mathrm{np}}(\rho_*, r_*)$ and cannot attain the efficiency bound. Since full-data covariate shift (DS.3) under the second parametrization in the above lemma is a special case of Condition DS.0† under the common parametrization $(P_A, P_{Z_1|A}, \ldots, P_{Z_K|\bar{Z}_{K-1},A})$ of distributions P (Li and Luedtke (2023)), we also conclude that there is generally no regular and fully robust asymptotically linear estimator of r_* that can attain the efficiency bound for Condition DS.0† under this parametrization.

We ran a simulation similar to that in Section 4.3. The results, which are presented in Supplement S4.1, demonstrate our theoretical results about the efficiency gains and the asymptotic behavior of our proposed estimator $\hat{r}_{\mathrm{cov}}$.

6. Discussion.

We have developed a general framework for risk estimation under dataset shift. It will be important to understand how our methods can applied in problems such as model training. Due to space limitation, we defer a more detailed discussion to the Supplementary Material.

Funding. This work was supported in part by NSF Grant DMS 2046874, NIH Grants R01AI27271, R01CA222147, R01AG065276, R01GM139926, and Analytics at Wharton.

The Africa Health Research Institute's Demographic Surveillance Information System and Population Intervention Programme is funded by the Wellcome Trust (201433/Z/16/Z) and the South Africa Population Research Infrastructure Network (funded by the South African Department of Science and Technology and hosted by the South African Medical Research Council).

SUPPLEMENTARY MATERIAL

Supplementary Material to "Efficient and multiply robust risk estimation under general forms of dataset shift" (DOI: 10.1214/24-AOS2422SUPP; .pdf). Additional new results and proofs of theoretical results.

REFERENCES

ANGELOPOULOS, A. N., BATES, S., CANDÈS, E. J., JORDAN, M. I. and LEI, L. (2021). Learn then test: Calibrating predictive algorithms to achieve risk control. Preprint. Available at arXiv:2110.01052v5.

ANGELOPOULOS, A. N., BATES, S., FANNJIANG, C., JORDAN, M. I. and ZRNIC, T. (2023). Prediction-powered inference. Preprint. Available at arXiv:2301.09633v1.

ANGRIST, J. D. and KRUEGER, A. B. (1992). The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *J. Amer. Statist. Assoc.* **87** 328–336.

AZRIEL, D., BROWN, L. D., SKLAR, M., BERK, R., BUJA, A. and ZHAO, L. (2021). Semi-supervised linear regression. *J. Amer. Statist. Assoc.* **117** 2238–2251.

BALAKRISHNAN, S., KENNEDY, E. H. and WASSERMAN, L. (2023). The fundamental limits of structure-agnostic functional estimation. Preprint. Available at arXiv:2305.04116v1.

BASTANI, H. (2021). Predicting with proxies: Transfer learning in high dimension. *Manage. Sci.* **67** 2964–2984.

BENKESER, D. and VAN DER LAAN, M. (2016). The highly adaptive lasso estimator. In 2016 *IEEE International Conference on Data Science and Advanced Analytics* (*DSAA*) 689–696. IEEE.

BHATTACHARYA, P. K., CHERNOFF, H. and YANG, S. S. (2007). Nonparametric estimation of the slope of a truncated regression. *Ann. Statist.* **11** 505–514. https://doi.org/10.1214/aos/1176346157

BICKEL, P., KLAASSEN, C. A., RITOV, Y. and WELLNER, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press, Baltimore, MD.

BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671.

BOLTHAUSEN, E., PERKINS, E. and VAN DER VAART, A. (2002). *Lectures on Probability Theory and Statistics*: *Ecole D'Eté de Probabilités de Saint-Flour XXIX*-1999. *Lecture Notes in Math.* **1781**. Springer, Berlin, Heidelberg.

BROOKHART, M. A. and VAN DER LAAN, M. J. (2006). A semiparametric model selection criterion with applications to the marginal structural model. *Comput. Statist. Data Anal.* **50** 475–498.

CAI, T., LI, M. and LIU, M. (2022). Semi-supervised triply robust inductive transfer learning. Preprint. Available at arXiv:2209.04977.

CAI, T. T. and WEI, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *Ann. Statist.* **49** 100–128.

CHAKRABORTTY, A. and CAI, T. (2018). Efficient and adaptive linear regression in semi-supervised settings. *Ann. Statist.* **46** 1541–1572.

CHATTERJEE, N., CHEN, Y. H., MAAS, P. and CARROLL, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *J. Amer. Statist. Assoc.* **111** 107–117.

CHEN, T. and GUESTRIN, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13–17 *August* 2016 785–794.

CHEN, X. and POUZO, D. (2015). Sieve Wald and QLR inferences on semi/nonparametric conditional moment models. *Econometrica* **83** 1013–1079.

CHEN, Y. H. and CHEN, H. (2000). A unified approach to regression analysis under double-sampling designs. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 449–460.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C. and NEWEY, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *Amer. Econ. Rev.* **107** 261–265.

CHERNOZHUKOV, V., CHETVERIKOV, D., DEMIRER, M., DUFLO, E., HANSEN, C., NEWEY, W. and ROBINS, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econom. J.* **21** C1–C68.

CHRISTODOULIDIS, S., ANTHIMOPOULOS, M., EBNER, L., CHRISTE, A. and MOUGIAKAKOU, S. (2017). Multisource transfer learning with convolutional neural networks for lung pattern analysis. *IEEE J. Biomed. Health Inform.* **21** 76–84. https://doi.org/10.1109/JBHI.2016.2636929

D'ORAZIO, M., DI ZIO, M. and SCANU, M. (2006). *Statistical Matching*: *Theory and Practice*. Wiley, Hoboken, NJ.

D'ORAZIO, M., DI ZIO, M. and SCANU, M. (2010). Old and new approaches in statistical matching when samples are drawn with complex survey designs. In *Proceedings of the* 45*th "Riunione Scientifica della Societa'Italiana di Statistica"*, *Padova* 16–18.

DORAN, G., MUANDET, K., ZHANG, K. and SCHÖLKOPF, B. (2014). A permutation-based kernel conditional independence test. In *Uncertainty in Artificial Intelligence - Proceedings of the* 30*th Conference*, *UAI* 2014 132–141.

EVANS, K., SUN, B. L., ROBINS, J. and TCHETGEN TCHETGEN, E. J. (2021). Doubly robust regression analysis for data fusion. *Statist. Sinica* **31** 1285–1307.

FOSTER, D. J. and SYRGKANIS, V. (2023). Orthogonal statistical learning. *Ann. Statist.* **51** 879–908.

FRIEDMAN, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** 1189–1232.

FRIEDMAN, J. H. (2002). Stochastic gradient boosting. *Comput. Statist. Data Anal.* **38** 367–378.

GRONSBELL, J., LIU, M., TIAN, L. and CAI, T. (2022). Efficient evaluation of prediction rules in semi-supervised settings under stratified sampling. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 1353–1391.

GU, T., HAN, Y. and DUAN, R. (2022). Robust angle-based transfer learning in high dimensions. Preprint. Available at arXiv:2210.12759.

GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. *Springer Series in Statistics*. Springer New York, New York, NY. https://doi.org/10.1007/B97848

HAJEK, J. (1962). Asymptotically most powerful rank-order tests. *Ann. Math. Stat.* **33** 1124–1147.

HASTIE, T., BUJA, A. and TIBSHIRANI, R. (1995). Penalized discriminant analysis. *Ann. Statist.* **23** 73–102.

HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. CRC Press/CRC, Boca Raton, FL.

HAUSMAN, J. A. (1978). Specification tests in econometrics. *Econometrica* **46** 1251–1271.

HE, Y., LI, Q., HU, Q. and LIU, L. (2022). Transfer learning in high-dimensional semiparametric graphical models with application to brain connectivity analysis. *Stat. Med.* **41** 4112–4129.

HU, X. and LEI, J. (2023). A two-sample conditional distribution test using conformal prediction and weighted rank sum. *J. Amer. Statist. Assoc.* 1–19.

JEWELL, N. P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika* **72** 11–21. https://doi.org/10.1093/biomet/72.1.11

KENNEDY, E. H. (2020). Towards optimal doubly robust estimation of heterogeneous causal effects. Preprint. Available at arXiv:2004.14497v3.

KOUW, W. M. and LOOG, M. (2018). An introduction to domain adaptation and transfer learning. Preprint. Available at arXiv:1812.11806.

LI, S., CAI, T. T. and LI, H. (2022). Transfer learning in large-scale Gaussian graphical models with false discovery rate control. *J. Amer. Statist. Assoc.* 1–13.

LI, S., GILBERT, P. B. and LUEDTKE, A. (2023). Data fusion using weakly aligned sources. Preprint. Available at arXiv:2308.14836v1.

LI, S. and LUEDTKE, A. (2023). Efficient estimation under data fusion. *Biometrika* **110** 1041–1054. https://doi.org/10.1093/biomet/asad007

LIPTON, Z., WANG, Y.-X. and SMOLA, A. (2018). Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning* 3122–3130. PMLR.

LIU, L., MUKHERJEE, R. and ROBINS, J. M. (2020). On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statist. Sci.* **35** 518–539.

LIU, L., MUKHERJEE, R. and ROBINS, J. M. (2023). Can we falsify the justification of the validity of Wald confidence intervals of doubly robust functionals, without assumptions? Preprint. Available at arXiv:2306.10590v1.

LIU, M., ZHANG, Y. and CAI, T. (2020). Augmented transfer regression learning with semi-non-parametric nuisance models. Preprint. Available at arXiv:2010.02521.

LIU, Y., LIU, M., GUO, Z. and CAI, T. (2023). Surrogate-assisted federated learning of high dimensional electronic health record data. Preprint. Available at arXiv:2302.04970v1.

MA, C., PATHAK, R. and WAINWRIGHT, M. J. (2023). Optimally tackling covariate shift in Rkhs-based nonparametric regression. *Ann. Statist.* **51** 738–761. https://doi.org/10.1214/23-AOS2268

MANSKI, C. F. and LERMAN, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica* **45** 1977.

MASON, L., BAXTER, J., BARTLETT, P. L. and FREAN, M. (1999). Boosting algorithms as gradient descent. In *Advances in Neural Information Processing Systems* **12**.

MCGRATH, S. and MUKHERJEE, R. (2022). On undersmoothing and sample splitting for estimating a doubly robust functional. Preprint. Available at arXiv:2212.14857v1.

MORENO-TORRES, J. G., RAEDER, T., ALAIZ-RODRÍGUEZ, R., CHAWLA, N. V. and HERRERA, F. (2012). A unifying view on dataset shift in classification. *Pattern Recognit.* **45** 521–530.

NEWEY, W. K. (1985). Generalized method of moments specification testing. *J. Econometrics* **29** 229–256.

NEWEY, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* **62** 1349.

NIE, X. and WAGER, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* **108** 299–319.

PAN, S. J. and YANG, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22** 1345–1359.

PARK, S., DOBRIBAN, E., LEE, I. and BASTANI, O. (2022). PAC prediction sets under covariate shift. In *International Conference on Learning Representations*.

PARK, S., LI, S., LEE, I. and BASTANI, O. (2020). PAC confidence predictions for deep neural network classifiers. Preprint. Available at arXiv:2011.00716.

PATHAK, R., MA, C. and WAINWRIGHT, M. J. (2022). A new similarity measure for covariate shift with applications to nonparametric regression. In *Proceedings of Machine Learning Research* **162** 17517–17530. PMLR.

PFANZAGL, J. (1985). *Contributions to a General Asymptotic Statistical Theory*. *Lecture Notes in Statistics* **3**. Springer New York, New York, NY.

PFANZAGL, J. (1990). *Estimation in Semiparametric Models*. *Lecture Notes in Statistics* **63**. Springer, New York, NY.

POLO, F. M., IZBICKI, R., LACERDA, E. G., IBIETA-JIMENEZ, J. P. and VICENTE, R. (2022). A unified framework for dataset shift diagnostics. Preprint. Available at arXiv:2205.08340.

QIU, H., DOBRIBAN, E. and TCHETGEN TCHETGEN, E. (2022). Prediction sets adaptive to unknown covariate shift. Preprint. Available at arXiv:2203.06126v5.

QIU, H., TCHETGEN TCHETGEN, E. and DOBRIBAN, E. (2024). Supplement to "Efficient and multiply robust risk estimation under general forms of dataset shift." https://doi.org/10.1214/24-AOS2422SUPP

RÄSSLER, S. (2012). *Statistical Matching*: *A Frequentist Theory*, *Practical Applications*, *and Alternative Bayesian Approaches*. *Lecture Notes in Statistics* **168**. Springer, New York, NY.

ROBINS, J. M., HSIEH, F. and NEWEY, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *J. Roy. Statist. Soc. Ser. B* **57** 409–424.

ROTNITZKY, A., FARAGGI, D. and SCHISTERMAN, E. (2006). Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. *J. Amer. Statist. Assoc.* **101** 1276–1288.

RUBIN, D. and VAN DER LAAN, M. J. (2007). A doubly robust censoring unbiased transformation. *Int. J. Biostat.* **3** Article 4. https://doi.org/10.2202/1557-4679.1052

SCHICK, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14** 1139–1151.

SCHÖLKOPF, B., JANZING, D., PETERS, J., SGOURITSA, E., ZHANG, K. and MOOIJ, J. M. (2012). On causal and anticausal learning. In *ICML*.

SCOTT, C. (2019). A generalized Neyman-Pearson criterion for optimal domain adaptation. In *Proceedings of the* 30*th International Conference on Algorithmic Learning Theory* **98** 738–761. PMLR.

SHAH, R. D. and PETERS, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.* **48** 1514–1538.

SHIMODAIRA, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Statist. Plann. Inference* **90** 227–244.

STORKEY, A. (2013). When training and test sets are different: Characterizing learning transfer. In *Dataset Shift in Machine Learning* 2–28. MIT Press, Cambridge.

SUGIYAMA, M. and KAWANABE, M. (2012). *Machine Learning in Non-stationary Environments*: *Introduction to Covariate Shift Adaptation*. MIT Press, Cambridge, MA.

SUGIYAMA, M., KRAULEDAT, M. and MULLER, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.* **8** 985–1005.

TASCHE, D. (2017). Fisher consistency for prior probability shift. *J. Mach. Learn. Res.* **18** 1–32.

TCHETGEN TCHETGEN, E. J. (2009). A commentary on G. Molenberghs's review of missing data methods. *Drug Inf. J.* **43** 433–435.

TIAN, Y. and FENG, Y. (2022). Transfer learning under high-dimensional generalized linear models. *J. Amer. Statist. Assoc.* 1–14.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288.

VANSTEELANDT, S. and DUKES, O. (2022). Assumption-lean inference for generalised linear model parameters. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **84** 657–685.

VANSTEELANDT, S., ROTNITZKY, A. and ROBINS, J. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika* **94** 841–860. https://doi.org/10.1093/biomet/asm070

VAN DER LAAN, M. (2017). A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *Int. J. Biostat.* **13**. https://doi.org/10.1515/ijb-2015-0097

VAN DER LAAN, M. J., POLLEY, E. C. and HUBBARD, A. E. (2007). Super learner. *Stat. Appl. Genet. Mol. Biol.* **6** Article 25. https://doi.org/10.2202/1544-6115.1309

VAN DER LAAN, M. J. and ROSE, S. (2018). *Targeted Learning in Data Science*: *Causal Inference for Complex Longitudinal Studies*. Springer, New York, NY.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. Cambridge Univ. Press, Cambridge, UK.

VAN DER VAART, A. W., DUDOIT, S. and VAN DER LAAN, M. J. (2006). Oracle inequalities for multi-fold cross validation. *Statist. Decisions* **24** 351–371.

VAN DER VAART, A. W. and WELLNER, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York, NY.

VAPNIK, V. (1992). Principles of risk minimization for learning theory. *Adv. Neural Inf. Process. Syst.* **4** 831–838.

VOVK, V. (2013). Conditional validity of inductive conformal predictors. *Mach. Learn.* **92** 349–376.

YANG, Y., KUCHIBHOTLA, A. K. and TCHETGEN TCHETGEN, E. (2024). Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **86** 943–965. https://doi.org/10.1093/jrsssb/qkae009

YUVAL, O. and ROSSET, S. (2023). Mixed semi-supervised generalized-linear-regression with applications to deep learning. Preprint. Available at arXiv:2302.09526v1.

ZHANG, A., BROWN, L. D. and CAI, T. T. (2019). Semi-supervised inference: General theory and estimation of means. *Ann. Statist.* **47** 2538–2566.

ZHANG, K., PETERS, J., JANZING, D. and SCHÖLKOPF, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the* 27*th Conference on Uncertainty in Artificial Intelligence, UAI* 2011 804–813. AUAI Press.

ZHANG, K., SCHÖLKOPF, B., MUANDET, K. and WANG, Z. (2013). Domain adaptation under target and conditional shift. In *International Conference on Machine Learning* 819–827. PMLR.

ZHANG, X., BLANCHET, J., GHOSH, S. and SQUILLANTE, M. S. (2022). A class of geometric structures in transfer learning: Minimax bounds and optimality. In *International Conference on Artificial Intelligence and Statistics* 3794–3820. PMLR.

ZHANG, Y. and BRADIC, J. (2022). High-dimensional semi-supervised learning: In search of optimal inference of the mean. *Biometrika* **109** 387–403.

ZHANG, Y., CHAKRABORTTY, A. and BRADIC, J. (2021). Double robust semi-supervised inference for the mean: Selection bias under mar labeling with decaying overlap. Preprint. Available at arXiv:2104.06667.

ZHOU, D., LIU, M., LI, M. and CAI, T. (2022). Doubly robust augmented model accuracy transfer inference with high dimensional features. Preprint. Available at arXiv:2208.05134.