Theory of Reweighting for Domain Adaptation and Causal Inference

Anonymous Author
Anonymous Institution

Abstract

Many approaches in domain adaptation and observational causal inference re-weight outcomes from a source population to estimate missing outcomes in a target population. We study weighting estimators, known as minimax weights, that minimize the worst-case error under restrictions on the outcome model. Under these restrictions, we derive a dual characterization by writing the variance of the weights as a ϕ-divergence. This characterization shows that the minimax weights are a (rescaled and recentered) member of the assumed outcome function class. We draw connections to density ratio estimation, but show that the source and target distribution do not need common support. Finally, we show conditions under which our results remain robust when our assumptions on the outcomes are wrong.

1 Introduction

Using covariates to transfer outcome information from one setting to another is a central task in science and engineering, including for domain adaptation, observational causal inference, and missing data imputation. These tasks share a common structure: we observe covariates and outcomes for a source distribution and want to predict outcomes given covariates in a target data set, which might have a different covariate distribution than the source. One standard approach is to reweight the source distribution to have a similar covariate distribution to the target, known as *importance weighting* for domain adaptation under covariate shift (Sugiyama et al., 2007a) and *inverse propensity score weighting* (IPW) for observational causal inference (Rosenbaum and Rubin, 1983). When the source and target distributions have common support, using the density ratio for weights leads to unbiased estimation. Importantly, this approach does not require any restrictions on the outcome model.

This approach has several drawbacks, however. First, using the density ratio for weights can lead to extremely large variance and unstable estimation (Kang et al., 2007; Cortes et al., 2010). Second, the density ratio is notoriously difficult to estimate, and simple plug-in estimates do not guarantee covariate balance between the reweighted source distribution and target distributions (see Ben-Michael et al., 2021).

An alternative approach instead finds weights that directly target covariate balance, with the choice of balance measure motivated (often implicitly) by restrictions on the outcome model (e.g., Gretton et al., 2009; Imai and Ratkovic, 2014). Of particular interest are so-called *minimax balancing weights*, which constrain the worst-case bias between groups over an outcome function class, enabling a bias-variance trade-off (see Zubizarreta, 2015; Hirshberg et al., 2019; Kallus, 2020).

1.1 Contributions

We begin with the premise that restrictions on the outcome model, commonly used in the literature, are typically necessary in practice. In particular, without any restrictions, the only weights that guarantee finite bias are the density ratio, which may have extreme variance or may not exist.

Given such restrictions, we then provide a new dual characterization of the minimax weights. We do so by interpreting the variance of the weights as a χ^2 divergence and then applying variational representations of ϕ-divergences. We use this characterization to make two points. First, we show that the minimax weights are always a (rescaled and recentered) function from the assumed outcome function class. Thus, if the outcome is assumed to be in an RKHS, the corresponding weights will be in the same RKHS. We precisely char-

Preliminary work. Under review by AISTATS 2022. Do not distribute.

acterize this relationship and also show computational gains from this alternative approach.

Second, we show that after making an assumption on the outcome function we no longer need the density ratio to exist. We connect the dual problem to a generalized form of density ratio estimation, *without* the requirement that the source and target distribution have common support. While it is well known that outcome modeling strategies do not require common support, this result is somewhat surprising for weighting estimators, where this assumption is common (however, see Kallus, 2020).

Finally, given the central role of restrictions on the outcome model, we briefly consider the setting in which this assumption is incorrect. In particular, we provide simple moment conditions under which we can retain a finite bound on the error when the true outcome model is not in the assumed class.

1.2 Related work

Estimators that target balance. Many reweighting estimators in causal inference explicitly target the discrepancy between source and target distribution (called *balance*). Examples include Hainmueller (2012); Zubizarreta (2015); Athey et al. (2018); Hirshberg et al. (2019); Tan (2020). See Ben-Michael et al. (2021) for a summary. The literature on domain adaptation also uses worst-case discrepancy between distributions (Mansour et al., 2009; Gretton et al., 2009; Yu and Szepesvári, 2012; Courty et al., 2014). Some approaches learn representations that minimize these discrepancies (Ganin et al., 2016; Shen et al., 2018; Assaad et al., 2021). Closely related are estimators that directly estimate the density ratio between two groups through a surrogate loss (Sugiyama et al., 2007b; Nguyen et al., 2010; Sugiyama et al., 2012).

Overlap in causal inference and adversarial training. Many existing theoretical treatments of balancing weights require the density ratio to exist (called *overlap* in causal inference), which is typically used for proving asymptotic consistency (see Hirshberg et al., 2019; Kallus, 2020). This assumption, however, can be highly restrictive especially in high-dimensions, as illustrated by D'Amour et al. (2021). See also Khan and Tamer (2010) for a discussion of the implications of overlap violations for causal inference. The same topic arises in adversarial training. See for example, Dupuis and Mao (2019); Glaser et al. (2021); Birrell et al. (2020a), who generalize ϕ-divergences to distributions that do not have common support. This idea is applied to GANs in Song and Ermon (2020).[1]

Domain adaptation and causal inference. Recent work combines ideas from these two literatures. For example, Shalit et al. (2017); Johansson et al. (2020) use integral probability metrics to estimate causal effects without the need for an overlap assumption. The same idea is used in Kallus (2020) for matching estimators in causal inference. Other work has made the connection between causal inference and adversarial training (Yoon et al., 2018; Ozery-Flato et al., 2018).

2 Problem Setup

Let $X \in \mathcal{X}$ denote covariates, and $Y \in \mathbb{R}$ denote outcomes. We study the domain adaptation problem with source and target populations, P and Q, with different joint distributions over X and Y. We observe X in both populations, but only observe the outcomes, Y, for the source population, P. The goal is to estimate the mean outcome in the target population, $\mathbb{E}_Q[Y]$.

We consider reweighting estimators, which use the known differences in X to transfer information about Y from P to Q. A key assumption for reweighting estimators is the *covariate shift* or *ignorability* assumption, which requires the relationship between covariates and outcomes to be the same across the two groups:

Assumption 1 (Ignorability)**.** *For all $x \in \mathcal{X}$,*

$$P(Y|X = x) = Q(Y|X = x).$$

Given Assumption 1, we can estimate $\mathbb{E}_Q[Y]$ using $\mathbb{E}_P[w(X)Y]$ with weights w that are a function of the observed covariates. Typically, Assumption 1 is paired with a restriction that the density ratio dQ/dP exists, also known as *overlap* or *continuity* in different literatures:

Definition (Common Support)**.** *We say P and Q have* common support *if Q is absolutely continuous with respect to P.*

In the special case where P and Q have common support and Assumption 1 holds, then $w = dQ/dP$ results in an unbiased estimator:

$$\begin{aligned}\mathbb{E}_P\left[\frac{dQ}{dP}(X)\, Y\right] &= \mathbb{E}_P\left[\frac{dQ}{dP}(X)\, \mathbb{E}_P[Y|X]\right]\\ &= \mathbb{E}_Q[\mathbb{E}_P[Y|X]] = \mathbb{E}_Q[Y],\end{aligned}$$

where we use ignorability for the last equality. We will consider weights such that $\mathbb{E}_P[w(X)] = 1$, i.e. we always end up with the same "size" population that we started with. However, in general we will *not* assume that P and Q have common support.

[1] Equation (11) in Song and Ermon (2020) is exactly the same as the minimax balancing objective from causal inference, where r are the weights, f is the dispersion, and T corresponds to the functions to balance.

2.1 Mean Squared Error

We would like to choose weights $w(X)$ that minimize the mean squared error (MSE) of $\mathbb{E}_P[w(X)Y]$ for estimating $\mathbb{E}_Q[Y]$. Define the *outcome function*, $f_0(x) \coloneqq \mathbb{E}_P[Y|X=x] = \mathbb{E}_Q[Y|X=x]$ and likewise, let $\sigma_0^2(x)$ be the conditional variance of Y. We expand the MSE via the standard bias-variance decomposition:

$$\begin{aligned}
\mathrm{MSE}(w) &= \mathbb{E}_P[(w(X)Y - \mathbb{E}_Q[Y])^2] \\
&= (\mathbb{E}_P[w(X)Y] - \mathbb{E}_Q[Y])^2 + \mathrm{Var}_P[w(X)Y] \\
&= (\mathbb{E}_P[w(X)f_0(X)] - \mathbb{E}_Q[f_0(X)])^2 \qquad (1) \\
&\quad + \mathbb{E}_P[w(X)^2\sigma_0^2(X)]. \qquad (2)
\end{aligned}$$

The MSE depends on two quantities: (1) the imbalance of the mean of the outcome function f_0 between the reweighted source distribution and the target distribution; and (2) the variability of the weights under the source distribution which amplifies the noise in the outcomes. In practice, it is convenient to consider either the homoskedastic case, $\sigma_0^2(x) = \sigma^2, \forall x$, or to upper-bound (2) via $\sigma^2\mathbb{E}_P[w^2]$, where $\sigma^2 \coloneqq \sup_{x\in\mathcal{X}} \sigma_0^2(x)$ and where we assume $0 < \sigma^2 < \infty$. We can then replace (2) by $\sigma^2\mathrm{Var}_P[w]$ without affecting the minimizer over w.

A natural idea is to choose weights that trade-off bias and variance to minimize the MSE. While f_0 is unknown in practice, if we assume that f_0 belongs to some function class $\mathcal{F}$, we can upper bound (1) by the worst-case difference in means over $\mathcal{F}$. We can then find weights that minimize the resulting bound on the MSE, sometimes known as *minimax weights* in observational causal inference (Hirshberg et al., 2019).

2.2 Notation

We will proceed more formally. Let $(\mathcal{X}, \mathcal{S})$ be a measurable space.[2] Let P and Q be given probability measures on $(\mathcal{X}, \mathcal{S})$. Let f_0 be a real-valued measurable function on $\mathcal{X}$. Denote $\mathcal{M}(\mathcal{X})$ the space of signed finite measures on $(\mathcal{X}, \mathcal{S})$ and $\mathcal{M}(P)$ those absolutely continuous with respect to (a.c. w.r.t) P. Denote $\mathcal{P}(\mathcal{X})$ the space of probability measures on $(\mathcal{X}, \mathcal{S})$ and $\mathcal{P}(P)$ those a.c. w.r.t. P.

With a slight abuse of notation, for measurable $f : \mathcal{X} \to \mathbb{R}$ and *both* $M \in \mathcal{M}(\mathcal{X})$ and $M \in \mathcal{P}(\mathcal{X})$, we will write $\mathbb{E}_M[f] \coloneqq \int_{\mathcal{X}} f(x)dM(x)$. We assume $\mathbb{E}_P[|f_0|] < \infty$ and $\mathbb{E}_Q[|f_0|] < \infty$.

The problem of choosing weights can be reformulated as finding a measure $R \in \mathcal{M}(P)$ such that $\int_{\mathcal{X}} dR(x) = 1$, with $w \coloneqq dR/dP$. This corresponds to the intuition behind reweighting as creating a "pseudo-population" based on P intended to match Q. We will therefore often use w and R interchangably.

While our setting and results are quite general, it may be helpful for the reader to keep in mind the case where $\mathcal{X}$ is finite and discrete with cardinality n. In this case, P and Q are probability vectors of length n and measurable functions are simply vectors in $\mathbb{R}^n$. Likewise, $\mathcal{M}(\mathcal{X})$ is just $\mathbb{R}^n$.

2.3 Assumptions on the Outcome Function

A bias-variance trade-off only exists if we place restrictions on f_0. When f_0 is completely unrestricted, for any $w \neq dQ/dP$, there always exists an f_0 that can make the bias term (1) arbitrarily large. As we discuss above, setting $w = dQ/dP$ has several downsides, including possibly extreme weights. To make progress, we can instead assume that f_0 belongs to some function class $\mathcal{F}$:

Assumption 2. *The outcome function f_0 belongs to $\mathcal{F}$ where $\mathcal{F}$ is a closed and convex set of measurable real-valued functions such that for all $f \in \mathcal{F}$, $\mathbb{E}_P[|f|] < \infty, \mathbb{E}_Q[|f|] < \infty$, and $-f \in \mathcal{F}$.*

In practice, for casual inference and missing value imputation, Assumption 2 requires making an assumption about the relation of the outcome of interest to the covariates. For the covariate shift problem or for augmented estimators like Hirshberg and Wager (2017), Assumption 2 requires making an assumption about the relationship of the accuracy of a predictor to its input features.

Many choices of $\mathcal{F}$ in Assumption 2 are quite general and justifiable with domain knowledge. Some examples for $0 < B < \infty$ are:

$$\begin{aligned}
\text{Bounded functions: } &\mathcal{F}_\infty \coloneqq \{f : \|f\|_\infty \leq B\} \\
\text{Lipschitz functions: } &\mathcal{F}_{\mathrm{Lip(c)}} \coloneqq \{f : \|f\|_{\mathrm{Lip(c)}} \leq B\} \\
\text{RKHS functions: } &\mathcal{F}_{\mathcal{H}} \coloneqq \{f : \|f\|_{\mathcal{H}} \leq B\},
\end{aligned}$$

where $\|\cdot\|_{\mathrm{Lip(c)}}$ denotes the Lipschitz constant with respect to a metric c and $\|\cdot\|_{\mathcal{H}}$ denotes the norm in some Reproducing Kernel Hilbert Space (RKHS), $\mathcal{H}$.

Under Assumption 2, the bias is bounded by the worst-case discrepancy in means over $\mathcal{F}$. This quantity is an example of an *integral probability metric* (IPM), defined for any set of functions, $\mathcal{G}$, and any $M, N \in \mathcal{M}(\mathcal{X})$ as:[3]

$$\mathrm{IPM}_{\mathcal{G}}(M, N) \coloneqq \sup_{g\in\mathcal{G}} \big\{|\mathbb{E}_M[g] - \mathbb{E}_N[g]|\big\}.$$

The bias term (1) for a re-weighted population R under

[2] To side-step topological issues, we assume that $\mathcal{X}$ is a separable Banach space.

[3] If $g \in \mathcal{G} \implies -g \in \mathcal{G}$ then the absolute value can be omitted.

Assumption 2 is upper-bounded by:

$$|\mathbb{E}_Q[f_0] - \mathbb{E}_R[f_0]| \leq \mathrm{IPM}_{\mathcal{F}}(Q, R).$$

This value is always finite by our assumptions on $\mathcal{F}$ and we can trade it off against the variance of the weights.

We now define several quantities that will be useful in our discussion below, beginning with maximum and minimum bias.

Definition (Maximum and minimum bias)**.** *The maximum bias, $\delta_{\max}$, is the bias under uniform weights (when $R = P$). The minimum bias, $\delta_{\min}$, is the smallest bias achieveable by reweighting P.*

$$\delta_{\max} \coloneqq IPM_{\mathcal{F}}(Q, P) \tag{3}$$

$$\delta_{\min} \coloneqq \inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} \big\{ IPM_{\mathcal{F}}(Q, R) \big\}. \tag{4}$$

Since $R = P$ is feasible (4), $\delta_{\min} \leq \delta_{\max}$. If P and Q have common support, $R = Q$ is also feasible, which implies $\delta_{\min} = 0$.

Definition (Distribution-defining)**.** *$\mathcal{F}$ is* distribution-defining *if $\forall M, N \in \mathcal{P}(\mathcal{X})$, $IPM_{\mathcal{F}}(M, N) = 0$ if and only if $M = N$.*

For example, $\mathcal{F}_\infty$ and $\mathcal{F}_{\mathrm{Lip(c)}}$ are distribution-defining, as is $\mathcal{F}_{\mathcal{H}}$ for a universal kernel. When $\mathcal{F}$ is distribution-defining then *only* dQ/dP can achieve worst-case bias zero, $\delta_{\min} = 0$, and if Q and P do not have common support then $\delta_{\min} > 0$.

2.4 The Variance of the Weights

We now characterize the variance term in (2) as a special case of a class of information-theoretic divergences called ϕ-divergences. We use this to derive our duality result and to make a novel connection to work on density ratio estimation and adversarial training.

For any convex function ϕ with $\phi(1) = 0$, the *ϕ-divergence* between $M \in \mathcal{M}(\mathcal{X})$ and $N \in \mathcal{P}(\mathcal{X})$ is:

$$D_\phi(M||N) \coloneqq \mathbb{E}_N \left[\phi\left(dM/dN\right)\right],$$

where $D_\phi(M||N) = \infty$ if M and N do not have common support. Given this definition, the variance of the weights is exactly the divergence between R and P with $\phi(x) = x^2 - 1$. This is known as the χ^2 divergence, and we denote it $D_2(R||P)$.

$$\mathrm{Var}_P[w] = \mathbb{E}_P[w^2 - 1] = \mathbb{E}_P\left[\left(\frac{dR}{dP}\right)^2 - 1\right] = D_2(R||P).$$

We can then upper bound the variance term in (2) by $\sigma^2 D_2(R||P)$, where σ^2 is generally unknown and is regarded as a tuning parameter.

Variational representations. It is possible to express ϕ-divergences in a dual form, called *variational representations*, as a supremum over measurable functions. Let $M \in \mathcal{M}(\mathcal{X})$ and let $N \in \mathcal{P}(\mathcal{X})$. Let ϕ^* denote the convex conjugate of ϕ. Then Keziou (2003) and Nguyen et al. (2005) show that:

$$D_\phi(M||N) = \sup_f \big\{ \mathbb{E}_M[f] - \mathbb{E}_N[\phi^*(f)] \big\}, \tag{5}$$

where the supremum is over all real-valued measurable functions on $\mathcal{X}$. If we additionally assume, as we do for R, that $\mathbb{E}_M[1] = 1$, then we have the tighter representation,

$$D_\phi(M||N) = \sup_f \big\{ \mathbb{E}_M[f] - \Lambda_N^\phi[f] \big\} \tag{6}$$

$$\text{where } \Lambda_N^\phi[f] \coloneqq \inf_{\lambda \in \mathbb{R}} \{ \lambda + \mathbb{E}_N[\phi^*(f - \lambda)] \}.$$

This result, using the infimum over λ in the spirit of Ruderman et al. (2012), appears to have been independently proposed by Agrawal and Horel (2020) and Birrell et al. (2020b). Under minimal conditions on ϕ, the suprema in (5) and (6) are achieved by $\phi'(dM/dN)$. We will show that estimating the minimax weights given Assumption 2 is equivalent to maximizing (6) over the function class $\mathcal{F}$.

2.5 Minimax Optimal Weights

We can combine the upper bounds from Sections 2.3 and 2.4 to bound the worst-case MSE. We refer to the weights that minimize this bound as *minimax optimal weights*, or just minimax weights. These weights solve the following optimization problem:

$$\inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} \left\{ \sup_{f \in \mathcal{F}} \{ \mathbb{E}_Q[f] - \mathbb{E}_R[f] \}^2 + \sigma^2 \mathbb{E}_P\left[\left(\frac{dR}{dP}\right)^2 - 1\right] \right\} \tag{7}$$

$$= \inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} \big\{ \mathrm{IPM}_{\mathcal{F}}(Q, R)^2 + \sigma^2 D_2(R||P) \big\}$$

A solution always exists because the objective is finite for $R = P$, which is feasible. For $\sigma^2 > 0$, the problem is strongly convex in R and has a unique solution.

Furthermore, $\exists \delta > 0$ such that (7) has the same minimizer as:

$$\inf_{\substack{R \in \mathcal{M}(P) \\ \mathbb{E}_R[1]=1}} D_2(R||P) \tag{8}$$

$$\text{such that } \mathrm{IPM}_{\mathcal{F}}(Q, R) \leq \delta.$$

We view σ^2 and δ as exchangeable tuning parameters: σ^2 represents the importance of reducing the variance of the weights; δ represents the level of acceptable bias. For $\sigma^2 \in (0, \infty)$, the corresponding δ lies in $(\delta_{\min}, \delta_{\max})$.

3 Duality Theory for Minimax Weights

We begin by giving a dual characterization of the solutions, R^*, to problems (7) and (8) and the corresponding minimax weights $w^* = dR^*/dP$.

3.1 Minimax Weights over a Function Class

In this section, we derive a duality result that characterizes the minimax weights under Assumption 2 where $f_0 \in \mathcal{F}$. The following theorem shows that the minimax weights are always a rescaled and recentered member of the function class $\mathcal{F}$, where the particular function and scaling factor depend on the worst-case bias tuning parameter δ.

Theorem 3.1. *Under Assumptions 1 and 2, for $\delta > \delta_{\min}$, the optimization problem (4) has a unique solution,*

$$\frac{dR^*}{dP} = 1 + \left(\frac{\mathbb{E}_Q[f^*] - \mathbb{E}_P[f^*] - \delta}{Var_P[f^*]}\right)(f^* - \mathbb{E}_P[f^*]),$$

where, for a unique $\mu \geq 0$ corresponding to δ, f^* achieves the following supremum:

$$\sup_{f \in \mathcal{F}} \left\{\mathbb{E}_Q[f] - \mathbb{E}_P[f] - \frac{\mu}{4} Var_P[f]\right\}. \tag{9}$$

The resulting MSE is:

$$MSE(R^*) \leq \delta^2 + \sigma^2 \frac{(\mathbb{E}_Q[f^*] - \mathbb{E}_P[f^*] - \delta)^2}{Var_p[f^*]}. \tag{10}$$

The proof is in the Appendix. The key idea is that solving the supremum in the constraint of (8) is hard for arbitrary $\mathcal{F}$. Duality allows us to exchange the subproblem over function classes for a sub-problem involving the χ^2 divergence. Then we apply the variational representation (6) to arrive at a single optimization problem over functions, (9).

Remark 3.1 (Other ϕ-Divergences)**.** We can replace the χ^2 divergence in the balancing weight problems (7) and (8) with other ϕ divergences. A duality result corresponding to Theorem 3.1 will hold for any convex function ϕ such that $\phi(1) = 0$ with convex conjugate ϕ^* such that $\{\phi^* < \infty\} = \mathbb{R}$. See the Appendix for details. We can use this general formulation to derive results for, e.g., non-negative weights or the KL divergence.

Remark 3.2 (Tuning Parameters)**.** Just like we can treat δ as a tuning parameter in (8) in place of an unknown σ^2, we can treat μ as a tuning parameter instead and solve (9) directly. The corresponding weights are:

$$\frac{dR^*}{dP} = 1 + \frac{\mu}{2}(f^* - \mathbb{E}_P[f^*]) \tag{11}$$

and we can recover the corresponding δ via:

$$\delta = \mathbb{E}_Q[f^*] - \mathbb{E}_P[f^*] - \frac{\mu}{2}\mathrm{Var}_P[f^*].$$

Remark 3.3 (Complete Information)**.** The case where we know f_0 exactly is a special case of Theorem 3.1 for $\mathcal{F}$ equal to the convex hull of $\{f_0, -f_0\}$. Assume without loss of generality[4] that $\mathbb{E}_Q[f_0] \geq \mathbb{E}_P[f_0]$. Then:

$$\frac{dR^*}{dP} = 1 + \left(\frac{\mathbb{E}_Q[f_0] - \mathbb{E}_P[f_0] - \delta}{\mathrm{Var}_P[f_0]}\right)(f_0 - \mathbb{E}_P[f_0]).$$

In this case, the optimal weights are always a rescaled and recentered version of f_0. The shape of weights does not depend on P, Q, or δ; only the scaling factor does. The MSE bound (10) becomes a quadratic in δ and we can solve for the optimal bias:

$$\delta^* = \left(\frac{\sigma^2}{\mathrm{Var}_P[f_0] + \sigma^2}\right)(\mathbb{E}_Q[f_0] - \mathbb{E}_P[f_0]),$$

which gives

$$\mathrm{MSE}(w^*) \leq \left(\frac{\sigma^2}{\mathrm{Var}_P[f_0] + \sigma^2}\right)(\mathbb{E}_Q[f_0] - \mathbb{E}_P[f_0])^2.$$

In other words, with complete information, we can analytically find the optimal bias-variance trade-off. Under homoskedasticity, these weights have the smallest MSE over all possible w such that $\mathbb{E}_P[w] = 1$.

3.2 The Function f^* Interpolates Between Two Extremes

We now show that as δ goes from $\delta_{\max}$ to $\delta_{\min}$, f^* interpolates smoothly between two extremes. The existing literature on balancing weights emphasizes the bias-variance trade-off, where weights with large bias are uniform, whereas weights with small bias can be highly dispersed. Using the dual optimal function f^*, we show how the *shape* of the weights changes as well.

For simplicity, we will discuss the setting where Q and P have common support and $\mathcal{F}$ is distribution-defining, so that $\delta_{\min} = 0$. We will return to the general setting in Section 4.

When $\delta = \delta_{\max}$, we have corresponding dual parameter $\mu = 0$. The solution to (9) is $f^* = f_{\max}$ which achieves the supremum for $\mathrm{IPM}(Q, P)$. In other words, $f_{\max}$ is

[4] Otherwise, replace f_0 with $-f_0$.

the function with worst-case bias between Q and P. At this extreme, plugging $\mu = 0$ into (11), the weights are uniform, and $R^* = P$.

At the second extreme, $\delta = 0$. By the distribution defining assumption, the weights equal dQ/dP, and $R^* = Q$. If $\mathcal{F}$ contains a rescaled/recentered version of the density ratio, then $\exists \mu_{\max} < \infty$ such that $f^* = f_{\text{ratio}}$, where:

$$f_{\text{ratio}} - \mathbb{E}_P[f_{\text{ratio}}] = \frac{2}{\mu_{\max}} \left(\frac{dQ}{dP} - 1 \right)$$

As the bias goes from $\delta_{\max}$ to 0, the shape of the weights, f^*, interpolates between $f_{\max}$ and f_{ratio}, scaled by the factor μ which goes from 0 to $\mu_{\max}$.

Example 1 (Discrete)**.** For a concrete example, consider discrete $\mathcal{X}$ with $|\mathcal{X}| = 3$. We choose $P = [0.2, 0.3, 0.5]$ and $Q = [0.5, 0.45, 0.05]$. We let $\mathcal{F} = \{f \in \mathbb{R}^3 : \|f\|_\infty \leq 1\}$. The shape of the minimax weights, as determined by the solution to (9), are extremely simple. For $\mu \in [0, 1]$, $f^* = f_{\max} = [1, 1, -1]$. For $\mu = \mu_{\max}$, $f^* = f_{\text{ratio}}$ is just the density ratio rescaled to fit in $[-1, 1]$. For $\mu \in (1, \mu_{\max})$, f^* is exactly all convex combinations of $f_{\max}$ and f_{ratio}. See Figure 1.

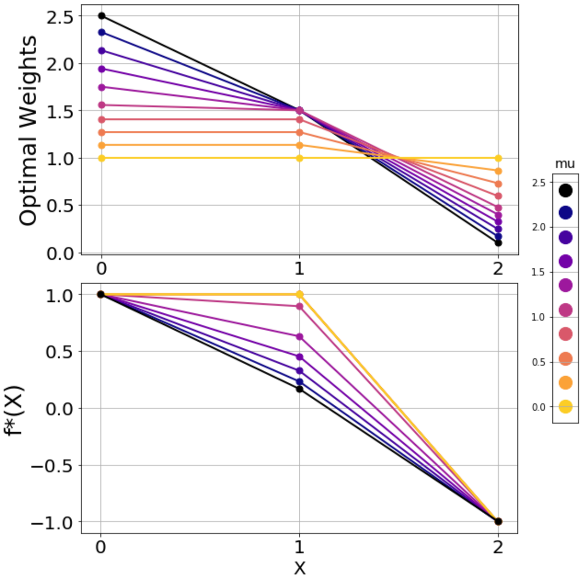

Figure 1: The optimal weights and the corresponding dual optimal function for the discrete example for μ ranging from 0 (yellow) to $\mu_{\max}$ (black).

Next, we give an example where $\mathcal{F}$ does not contain a rescaled version of dQ/dP:

Example 2 (Gaussian)**.** Let $\mathcal{X} = \mathbb{R}$. Let P be Gaussian with mean 1 and variance 1, let Q be Gaussian with mean 2 and variance 1, and let p and q denote their densities. Let $\mathcal{F} = \{f : \|f\|_\infty \leq 1\}$.

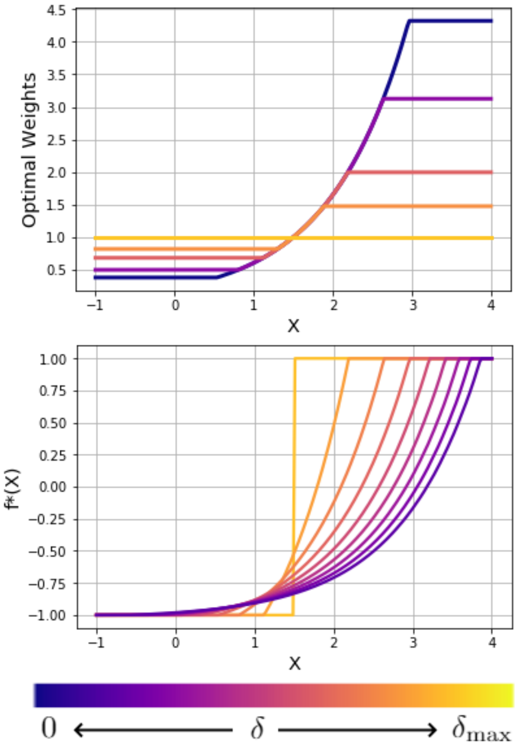

Figure 2: The optimal weights and corresponding dual optimal function for the Gaussian example for δ starting at $\delta_{\max}$ and shrinking towards zero.

On one extreme, for $\mu = 0$, $f_{\max}$ has an explicit form: $f_{\max}(x) = 1$ when $q(x) \geq p(x)$ and $f_{\max}(x) = -1$ when $q(x) < p(x)$.

On the other hand, the density ratio is not bounded, and so there is no f_{ratio} which is a rescaled version of q/p in $\mathcal{F}$ and no $\mu_{\max} < \infty$. Instead, as $\delta \to 0$ the corresponding $\mu \to \infty$.

As a result, the minimax weights still have a simple characterization. As μ goes from $0 \to \infty$, f^* starts at $f_{\max}$ and becomes shaped more and more like q/p but constrained to $[-1, 1]$. See the bottom panel of Figure 2. The corresponding δ goes from $\delta_{\max} \to 0$ and for each δ the minimax weights are exactly equal to q/p but truncated above and below. The level of truncation expands as μ increases, and the weights converge to the true density ratio. See the top panel of Figure 2.

4 Relationship to Density Ratio Estimation

The previous examples establish a clear connection to density ratio estimation. In this section, we use the dual problem (9) to show that the minimax weights problem is a generalized form of density ratio estimation, but which does not require common support once we have made Assumption 2.

We begin by establishing a clearer link between (9) and our characterization of the variance of the weights as a ϕ-divergence as used in the proof of Theorem 3.1. Consider the variational representation (6). Specializing to $\phi(x) = \mu(x^2 - 1)$ for $\mu > 0$, we can write the weighted χ^2 divergence between Q and P as:

$$\frac{1}{\mu} D_2(Q||P) = \sup_f \Big\{ \mathbb{E}_Q[f] - \mathbb{E}_P[f] - \frac{\mu}{4} \mathrm{Var}_P[f] \Big\},$$

where the supremum is over all real-valued measurable functions. If Q and P have common support then the supremum is achieved by $2\mu(dQ/dP)$, otherwise $D_2(Q||P) = \infty$. The variational representation is identical to (9), except that the problem in (9) is restricted to functions in $\mathcal{F}$.

We immediately have the following corollary:

Corollary 4.1. *Under the conditions in Theorem 3.1,*

$$\frac{2}{\mu}\frac{dQ}{dP} \in \mathcal{F} \implies f^* = \frac{2}{\mu}\frac{dQ}{dP} \text{ and } \frac{dR^*}{dP} = \frac{dQ}{dP}.$$

If the density ratio exists and a scaled version belongs to the outcome function class, then it is minimax optimal to reweight so that there is zero bias.

There is an immediate connection to density ratio estimation using ϕ-divergences (Nguyen et al., 2010), which also maximizes a variational representation within a function class. Unlike that approach, however, we do not require that the density ratio belongs to $\mathcal{F}$ or even exists. If Q and P do not have common support, then the supremum in (9) is still finite when restricted to $\mathcal{F}$.

Instead, we can reinterpret the dual solution and corresponding minimax weights as "density ratio" estimation in a more general sense. The true density ratio transforms P into Q with zero discrepancy on any function, which is not possible when P and Q lack common support. The minimax weights, dR^*/dP, are a ratio that transfer our knowledge about $Y \sim P$ into knowledge about $Y \sim Q$, and we do not need to make any assumption about the form of this ratio. Instead our assumptions about the relationship between Y and X pin down the functional form for us via (11).

5 IHDP Example

In this section, we walk through an application to make the previous discussion on density ratio estimation more concrete. We use an RKHS for the outcome function, resulting in the KOM estimator from Kallus (2020). However, we solve the problem in the dual to provide an intuitive characterization of the minimax weights.

5.1 The IHDP Dataset and Setup

The Infant Health and Development Program (IHDP) data set is a standard observational causal inference benchmark from Hill (2011), based on data from a randomized control trial of an intensive home visiting and childcare intervention for low birth weight infants born in 1985. We consider a non-experimental subset of the original data with $n_0 = 608$ children assigned to control, $n_1 = 139$ children assigned to treatment, and $n = 747$ total children. For all children, we have a range of baseline covariates, including both categorical covariates, like the mother's educational attainment, and continuous covariates, like the child's birth weight. Our goal is to estimate the average outcome (a standardized test score) in the absence of the intensive intervention. We observe this outcome for the 608 control children, and want to re-weight these observations to estimate the missing mean for the 139 treated children.

To do so, we use an RKHS as a flexible but tractable functional form for f_0. In particular, we assume that $\mathcal{F} = \mathcal{F}_{\mathcal{H}}^B \coloneqq \{f : \|f\|_{\mathcal{H}} \le B\}$ for $B < \infty$, where $\mathcal{H}$ is the RKHS induced by the Gaussian kernel,

$$\mathcal{K}(x_1, x_2) = \exp\left(-\frac{1}{2}\|x_1 - x_2\|_2^2\right).$$

Define $K \in \mathbb{R}^{n \times n}$ with $K_{ij} = \mathcal{K}(X_i, X_j)$. Then for any $f \in \mathcal{F}$, there exists an $\alpha \in \mathbb{R}^n$ such that $\alpha^T K \alpha \le B$ and $f(X_j) = \sum_{i=1}^n \alpha_i K_{ij}, \forall j$.

Problem (9) can be written as a simple quadratic optimization problem over these vectors α. We find the minimax weights for many values of the tuning parameter μ. See the Appendix for details.

Note the role of the smoothness of the function class $\mathcal{F}_{\mathcal{H}}^B$. Intuitively, we think of, say, $X_i = 1.71$ and $X_j = 1.72$ as being close to each other. For arbitrary f, however, $f(1.71)$ might be vary large and $f(1.72)$ might be very small. The kernel $\mathcal{K}$ provides a formal sense in which data points are close to each other.

5.2 The Minimax Weights

As in Examples 1 and 2, the minimax weights interpolate between two extremes. At one extreme we have $\mu = 0$, which finds uniform weights that minimize the variance. The solution to (9) is then $f^* = f_{\max}$, i.e., the function in $\mathcal{F}_{\mathcal{H}}^B$ with the largest difference in means between Q and P, and with corresponding bias $\delta_{\max}$.

At the other extreme, finding the weights with minimum bias takes more care. Unlike Examples 1 and 2, we do not assume that P and Q have common support and therefore there are no weights that achieve zero bias. We instead find weights that achieve the smallest possible bias over $\mathcal{F}_{\mathcal{H}}^B$, $\delta_{\min}$, which corresponds to some tuning parameter $\mu = \mu_{\max} < \infty$. Finally, the dual optimal $f^* = f_{\min}$ for $\mu_{\max}$ gives us the shape of the weights with minimum bias. We can think of these weights as an approximate "density ratio"—these are the weights that transform P to look as much like Q as possible on $\mathcal{F}^B_{\mathcal{H}}$, even though Q and P have disjoint support.

As we go from $\delta_{\max}$ to $\delta_{\min}$, the weights interpolate between $f_{\max}$ and $f_{\min}$ scaled by a factor μ which increases from 0 to $\mu_{\max}$. See Figure 3.

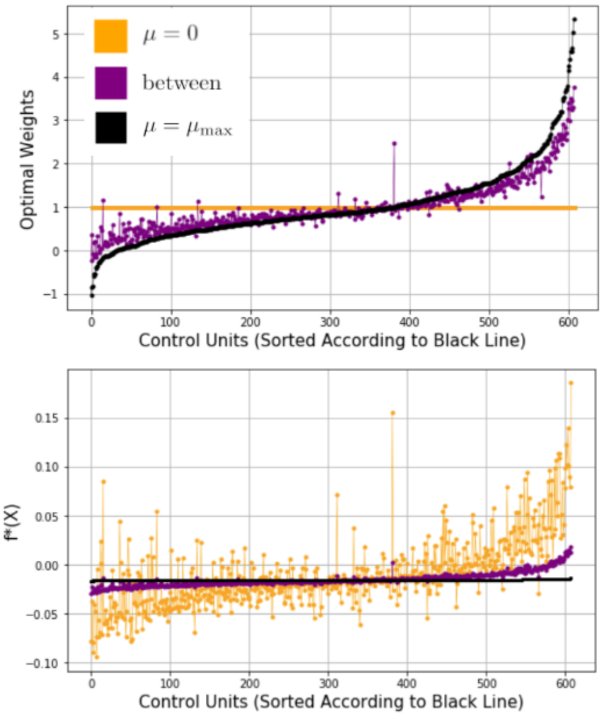

Figure 3: The optimal weights and corresponding dual optimal function for the IHDP example for the extreme values of μ and one intermediate value.

Remark 5.1 (The Role of the Kernel)**.** As the radius of the RKHS-norm ball B increases the value of $\mu_{\max}$ changes, but the weights that achieve $\delta_{\min}$ *remain the same.* Thus it is the smoothness properties of $\mathcal{K}$ that drive the qualitative behavior of the weights, not the value of the parameter B.

Remark 5.2 (Connection to Classification)**.** If we sort all data points (treated and control) according to the dual optimal solution f^* for $\mu = \mu_{\max}$, then the treated and control units are *perfectly sorted.*[5] This provides a direct connection to work on permutation weighting in Arbour et al. (2021) which solves the balancing weights problem by training a classifier.

Remark 5.3 (Computational Advantages of the Dual)**.** In some situations, it is computationally easier to solve the dual problem (9) directly instead of the primal problem (8). For an RKHS, there is a closed form of the IPM available which makes the primal and dual problems equally easy to solve. But consider a class of neural networks parameterized by network weights θ:

$$\mathcal{F}^B_{\text{NN}} \coloneqq \{f_\theta : \|\theta\| \leq B\}$$

Then handling the IPM constraint in the primal problem requires adversarial training which can be quite challenging. On the other hand, (9) requires training a neural network once with a convex loss function which can be accomplished with off-the-shelf SGD.

6 Robustness

Minimax weights rely heavily on the function class in Assumption 2. In this section, we show that with minimal moment conditions we can still retain a bound on the bias even if we have misspecified the function class $\mathcal{F}$. We consider two functions classes. First, $\mathcal{F}$ for which we solve (8) to find R^* such that $\text{IPM}_{\mathcal{F}}(Q, R^*) \leq \delta$. Second, the *true* function class, $\mathcal{G}$ such that $f_0 \in \mathcal{G}$ and $f_0 \notin \mathcal{F}$. To bound the bias, we need to show that

$$\text{IPM}_{\mathcal{F}}(Q, R^*) \leq \delta \implies \text{IPM}_{\mathcal{G}}(Q, R^*) \leq \rho(\delta) \quad (12)$$

for some $\rho < \infty$ which has good scaling with δ. Without further assumptions, (12) will *not* hold for any $\mathcal{G}$.

IPMs correspond to common perturbations in the robust statistics literature. For example, $\text{IPM}_{\mathcal{F}_\infty}$ and $\text{IPM}_{\mathcal{F}_{\text{Lip(c)}}}$ are equivalent to the total variation (TV) distance and Wasserstein distance respectively. For $\mathcal{F} = \mathcal{F}_\infty$, we can apply Lemma E.2 from Zhu et al. (2019) to achieve (12) for any $\mathcal{G}$. We require an Orlicz norm bound under Q and R^* on $g(X)$ for all $g \in \mathcal{G}$. For a simple example, let $\mathcal{G}$ be linear: $\{g = \beta^T x : \|\beta\| \leq 1\}$. Then we get the following two results:

Proposition 6.1. *If R^* and Q have bounded covariance, then we have the following upper bound on the bias:*

$$|\mathbb{E}_Q[f_0] - \mathbb{E}_{R^*}[f_0]| \leq \rho(\delta),$$

where $\rho(\delta) = O(\sqrt{\delta})$.

Proposition 6.2. *If R^* and Q are sub-Gaussian, then we have the following upper bound on the bias:*

$$|\mathbb{E}_Q[f_0] - \mathbb{E}_{R^*}[f_0]| \leq \rho(\delta),$$

where $\rho(\delta) = O(\delta\sqrt{\log(1/\delta)})$.

For general $\mathcal{G}$, the rate of ρ in terms of δ is similar, but the moment conditions on X become stronger. In practice, these robust statistics results mean that we can make a best guess about $\mathcal{F}$ but as long as Q is sufficiently "nice", the true bias will not be much larger than δ.

[5] This is not true for $\mu < \mu_{\max}$. In this case some higher weights are assigned to P and some lower weights are assigned to Q.

References

R. Agrawal and T. Horel. Optimal bounds between f-divergences and integral probability metrics. In *International Conference on Machine Learning*, pages 115–124. PMLR, 2020.

D. Arbour, D. Dimmery, and A. Sondhi. Permutation weighting. In *International Conference on Machine Learning*, pages 331–341. PMLR, 2021.

S. Assaad, S. Zeng, C. Tao, S. Datta, N. Mehta, R. Henao, F. Li, and L. Carin Duke. Counterfactual representation learning with balancing weights. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 1972–1980. PMLR, 13–15 Apr 2021. URL `https://proceedings.mlr.press/v130/assaad21a.html`.

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.

E. Ben-Michael, D. Hirshberg, A. Feller, and J. Zubizarreta. The balancing act in causal inference. 2021.

J. Birrell, P. Dupuis, M. A. Katsoulakis, Y. Pantazis, and L. Rey-Bellet. (f, γ)-divergences: Interpolating between f-divergences and integral probability metrics. *arXiv preprint arXiv:2011.05953*, 2020a.

J. Birrell, M. A. Katsoulakis, and Y. Pantazis. Optimizing variational representations of divergences and accelerating their statistical estimation. *arXiv preprint arXiv:2006.08781*, 2020b.

C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Nips*, volume 10, pages 442–450. Citeseer, 2010.

N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

P. Dupuis and Y. Mao. Formulation and properties of a divergence used to compare probability measures without absolute continuity. *arXiv preprint arXiv:1911.07422*, 2019.

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

Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

P. Glaser, M. Arbel, and A. Gretton. Kale flow: A relaxed kl gradient flow for probabilities with disjoint support. *arXiv preprint arXiv:2106.08929*, 2021.

A. Gretton, A. Smola, J. Huang, M. Schmittfull, K. Borgwardt, and B. Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

J. Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, pages 25–46, 2012.

J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

D. A. Hirshberg and S. Wager. Augmented minimax linear estimation. *arXiv preprint arXiv:1712.00038*, 2017.

D. A. Hirshberg, A. Maleki, and J. R. Zubizarreta. Minimax linear estimation of the retargeted mean. *arXiv preprint arXiv:1901.10296*, 2019.

K. Imai and M. Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

F. D. Johansson, U. Shalit, N. Kallus, and D. Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *arXiv preprint arXiv:2001.07426*, 2020.

N. Kallus. Generalized optimal matching methods for causal inference. *Journal of Machine Learning Research*, 21(62):1–54, 2020.

J. D. Kang, J. L. Schafer, et al. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4):523–539, 2007.

A. Keziou. Dual representation of φ-divergences and applications. *Comptes rendus mathématique*, 336 (10):857–862, 2003.

S. Khan and E. Tamer. Irregular identification, support conditions, and inverse weight estimation. *Econometrica*, 78(6):2021–2042, 2010.

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

X. Nguyen, M. J. Wainwright, and M. I. Jordan. On divergences, surrogate loss functions, and decentralized detection. *arXiv preprint math.ST/0510521*, 2005.

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

M. Ozery-Flato, P. Thodoroff, M. Ninio, M. Rosen-Zvi, and T. El-Hay. Adversarial balancing for causal inference. *arXiv preprint arXiv:1810.07406*, 2018.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

A. Ruderman, M. Reid, D. García-García, and J. Petterson. Tighter variational representations of f-divergences via restriction to probability measures. *arXiv preprint arXiv:1206.4664*, 2012.

U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

J. Shen, Y. Qu, W. Zhang, and Y. Yu. Wasserstein distance guided representation learning for domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

J. Song and S. Ermon. Bridging the gap between f-gans and wasserstein gans. In *International Conference on Machine Learning*, pages 9078–9087. PMLR, 2020.

M. Sugiyama, M. Krauledat, and K.-R. Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8 (5), 2007a.

M. Sugiyama, S. Nakajima, H. Kashima, P. Von Buenau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *NIPS*, volume 7, pages 1433–1440. Citeseer, 2007b.

M. Sugiyama, T. Suzuki, and T. Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012.

Z. Tan. Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107(1):137–158, 2020.

J. Yoon, J. Jordon, and M. Van Der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

Y. Yu and C. Szepesvári. Analysis of kernel mean matching under covariate shift. *arXiv preprint arXiv:1206.4650*, 2012.

B. Zhu, J. Jiao, and J. Steinhardt. Generalized resilience and robust statistics. *arXiv preprint arXiv:1909.08755*, 2019.

J. R. Zubizarreta. Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110 (511):910–922, 2015.