Learning Weighted Representations for Generalization Across Designs

Fredrik Johansson Nathan Kallus Uri Shalit David Sontag

October 22, 2020

Presented by: Serge Assaad

Basic Setup

- Dataset has covariate/ binary treatment/outcome triplets $\{x_i, t_i, y_i\}_{i=1}^{N}$.
- Latent potential outcomes $y_i(0), y_i(1)$ with $y_i = y_i(t_i)$ the observed outcome.
- Assume *ignorability* $(y_i(0), y_i(1) \perp t_i | x_i)$ and *positivity* $(e(x) \triangleq p(t = 1|x) > 0 \ \ \forall x)$ – together these are "strong ignorability."

Problem Statement & Motivation

- **Goal:** Predict outcomes on a "target design" $p_\pi(x,t)$ with m samples $(x'_1,t'_1),\ldots,(x'_m,t'_m)$, given n triplets $(x_1,t_1,y_1),\ldots,(x_n,t_n,y_n)$, where the pairs (x_i,t_i) are drawn from a "source design", $p_\mu(x,t)$.
- "design" is a general term the authors use for a joint distribution of covariates & treatment. A design $p(x,t)$ can be written as follows:

$$\underbrace{p(x,t)}_{\text{"design"}} = \underbrace{p(t|x)}_{\text{"policy"}} \cdot \underbrace{p(x)}_{\text{"domain"}}$$

- Motivating example: if all our observed data comes from a hospital P, but we want to predict counterfactuals for patients in a hospital Q, then we have to deal with an additional "domain shift". This rolls into one the problems of counterfactual prediction (from observed data) and domain adaptation.

Source vs. target design

- We can write source and target distributions for triplets (x, t, y) as:

$$\text{Target:} \quad p_\pi(x,t,y) = p_\pi(x)p_\pi(t|x)p(y|t,x) \tag{1}$$

$$\text{Source:} \quad p_\mu(x,t,y) = p_\mu(x)p_\mu(t|x)p(y|t,x) \tag{2}$$

- Note that it is assumed that the outcome generating process $p(y|x,t)$ is the same for both source and target – i.e., only the *design* $p(x,t)$ changes between source and target – this is similar to assumptions often made in domain adaptation.
- Again, note that we have samples $(x_i, t_i, y_i) \sim p_\mu(x,t,y)$ from the source, but we only have samples $x'_i, t'_i \sim p_\pi(x,t)$ from the target.

Reweighted risk & importance sampling

- Assume we have predictors $f : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ where $f(x, t)$ approximates the expected potential outcome $E[Y|X = x, T = t]$.
- We would like to quantify the quality of the predictor f on the target design $p_\pi(x, t)$. We do this with the target risk:

$$R_\pi(f) := \mathbb{E}_{x,t,y\sim p_\pi}[\ell_f(x, t, y)] \tag{3}$$

where l_f is a loss function (e.g., squared loss $(f(x, t) - y)^2$).

- The difficulty here is that we don't have samples y for target distribution, but we can rewrite the risk using *importance sampling* with weights $w^*(x, t) = p_\pi(x, t)/p_\mu(x, t)$, as follows:

$$R_\mu^{w^*}(f) := \mathbb{E}_{x,t,y\sim p_\mu}[w^*(x, t)\ell_f(x, t, y)] = R_\pi(f) \ . \tag{4}$$

Weighting in practice

- In practice, we don't actually have the importance sampling weights $w^*(x,t)$, but we may still define a reweighted density $p_\mu^w(x,t) := w(x,t)p_\mu(x,t)$ for some reweighting function $w(x,t)$, defined below:

Definition

A function $w : \mathcal{X} \times \mathcal{T} \to \mathbb{R}_+$ *is a valid re-weighting of* p_μ *if*

$$\mathbb{E}_{x,t\sim p_\mu}[w(x,t)] = 1 \;\; \textit{and} \;\; p_\mu(x,t) > 0 \Rightarrow w(x,t) > 0.$$

- The general approach of this paper is to *learn* a weighting function $w(x,t)$ from the observational data, as well as regressors $f(x,t)$

Bound on target risk

Lemma

For hypotheses f with loss ℓ_f such that $\ell_f / \|\ell_f\|_{\mathcal{H}} \in \mathcal{H}$, and p_μ, p_π with common support, there exists a valid re-weighting w, see Definition 1, such that,

$$\begin{aligned} R_\pi(f) &\leq R^w_\mu(f) + \|\ell_f\|_{\mathcal{H}} IPM_{\mathcal{H}}(p_\pi, p^w_\mu) \\ &\leq R_\mu(f) + \|\ell_f\|_{\mathcal{H}} IPM_{\mathcal{H}}(p_\pi, p_\mu) \ . \end{aligned} \tag{5}$$

The first inequality is tight for importance sampling weights, $w(x,t) = p_\pi(x,t)/p_\mu(x,t)$. The second inequality is not tight for general f, even if $\ell_f / \|\ell_f\|_{\mathcal{H}} \in \mathcal{H}$, unless $p_\pi = p_\mu$.

Bound on target risk

Theorem

Suppose that Φ is a twice-differentiable, invertible representation (with inverse Ψ), that $h(\Phi, t)$ is an hypothesis, and $f = h(\Phi(x), t) \in \mathcal{F}$. Define $m_t(x) = \mathbb{E}_Y[Y \mid X = x, T = t]$, let $\ell_{h,\Phi}(\Psi(z), t) := L(h(z, t), m_t(\Psi(z)))$ where L is the squared loss, $L(y, y') = (y - y')^2$, and assume that there exists a constant $B_\Phi > 0$ such that $\ell_{h,\Phi}/B_\Phi \in \mathcal{H} \subseteq \{h : \mathcal{Z} \times \mathcal{T} \to \mathcal{Y}\}$. Finally, let w be a valid re-weighting of $p_{\mu,\Phi}$. Then,

$$R_\pi(f) \leq R^w_\mu(f) + B_\Phi IPM_\mathcal{H}(p_{\pi,\Phi}, p^w_{\mu,\Phi}) + C \tag{6}$$

where C is a constant w.r.t. the parameters of h and Φ

ITE estimation as a special case of design shift

- The authors argue that ITE estimation is a special case within the framework of design shift.
- Namely, for ITE estimation, we have matching domains $p_\pi(x) = p_\mu(x)$.
- WLOG, suppose we would like to predict the potential outcome $Y(1)$.
  - The source is then the treatment group, with source policy $p_\mu(t|x) = \mathbb{1}(t = 1)$ ("treat-all" policy, denoted π_1)
  - the target is the control group, with target policy $p_\mu(t|x) = \mathbb{1}(t = 0)$ ("treat-none" policy, denoted π_0).

ITE estimation (cont.)

- To measure quality of ITE estimation, we use:

$$\text{MSE}(\hat{\tau}) = \mathbb{E}_p\left[(\hat{\tau}(x) - \tau(x))^2\right], \tag{7}$$

where $\tau(x) := E[Y|T=1, X=x] - E[Y|T=1, X=x]$ is the true ITE and $\hat{\tau}(x)$ is the estimated ITE (here, $\hat{\tau}(x) = f(x,1) - f(x,0)$).

Proposition

We have with MSE($\hat{\tau}$) as in (7) and $R_{\pi_t}(f)$ the risk under the constant policy π_t (t=1: "treat-all" or t=0: "treat-none"),

$$\textit{MSE}(\hat{\tau}) \leq 2(R_{\pi_1}(f) + R_{\pi_0}(f)) - 4\sigma^2 \tag{8}$$

where σ is such that $\forall t \in \mathcal{T}, x \in \mathcal{X}, \sigma_Y(x,t) \geq \sigma$ and $\sigma_Y^2(x,t)$ is variance of $Y(t)$ conditioned on $X = x$.

- We can use this proposition, combined with the previous risk bound, to get a practical upper bound for the MSE.

Training objective

- The authors use the following as the finite-sample approximation of their upper bound on target risk:

$$\mathcal{L}_\pi(h, \Phi, w; \beta) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} w_i \ell_h(\Phi(x_i), t_i) + \frac{\lambda_h}{\sqrt{n}}\mathcal{R}(h)}_{\mathcal{L}_\pi^h(h,\Phi,w;D,\alpha,\lambda_h)} + \underbrace{\alpha\ \mathrm{IPM}_G(\hat{p}_{\pi,\Phi}, \hat{p}^w_{\mu,\Phi}) + \lambda_w \frac{\|w\|_2}{n}}_{\mathcal{L}_\pi^w(\Phi,w;D,\alpha,\lambda_w)} \tag{9}$$

where $\mathcal{R}(h)$ is a regularizer of h, such as ℓ_2-regularization, and $\hat{p}_{\pi,\Phi}$, $\hat{p}^w_{\mu,\Phi}$ are empirical approximations of $p_{\pi,\Phi}, p^w_{\mu,\Phi}$, respectively.

Optimization steps

- Optimizing the objective presented is difficult in practice. As a heuristic, the authors train the weighting function $w(x, t)$ based only on the IPM imbalance term and L2-regularizer.
- They optimize the encoder Φ + regressor h, followed by the weighting function $w(x, t)$, in alternating fashion, as follows:

$$h^k, \Phi^k = \arg\min_{h,\Phi} \quad \mathcal{L}_\pi^h(h, \Phi, w; D, \alpha, \lambda_h), \tag{10}$$

$$w^{k+1} = \arg\min_{w} \quad \mathcal{L}_\pi^w(\Phi^k, w; D, \alpha, \lambda_w) \tag{11}$$

Synthetic experiment – domain adaptation

The authors formulate a toy problem, as follows:

- n labeled source samples, distributed according to $p_\mu(x) = \mathcal{N}(x; m_\mu, I_d)$, $m_\mu = \mathbf{1}_d/2$ (here, $d = 10$).
- n unlabeled target samples drawn according to $p_\pi(x) = \mathcal{N}(x; m_\pi, I_d)$, $m_\pi = -\mathbf{1}_d/2$.
- Source sample outcomes are generated as $y = \sigma(\beta^\top x + c)$, where $\beta \sim \mathcal{N}(\mathbf{0}_d, 1.5 I_d)$, $c \sim \mathcal{N}(0, 1)$.
- Note: Importance sampling weights $w^*(x) = p_\pi(x)/p_\mu(x)$ are known here.
- Use an identity representation $\Phi(x) = x$, and fit (misspecified) linear models $f(x) = \beta^\top x + \gamma$

They compare the following approaches:

- Parameterizing weights as a neural network (their proposal)
- Using importance sampling weights (known in closed form)
- Using clipped importance sampling weights $w_M(x) = \min(w(x), M)$, with $M \in \{5, 10\}$

Synthetic experiment results

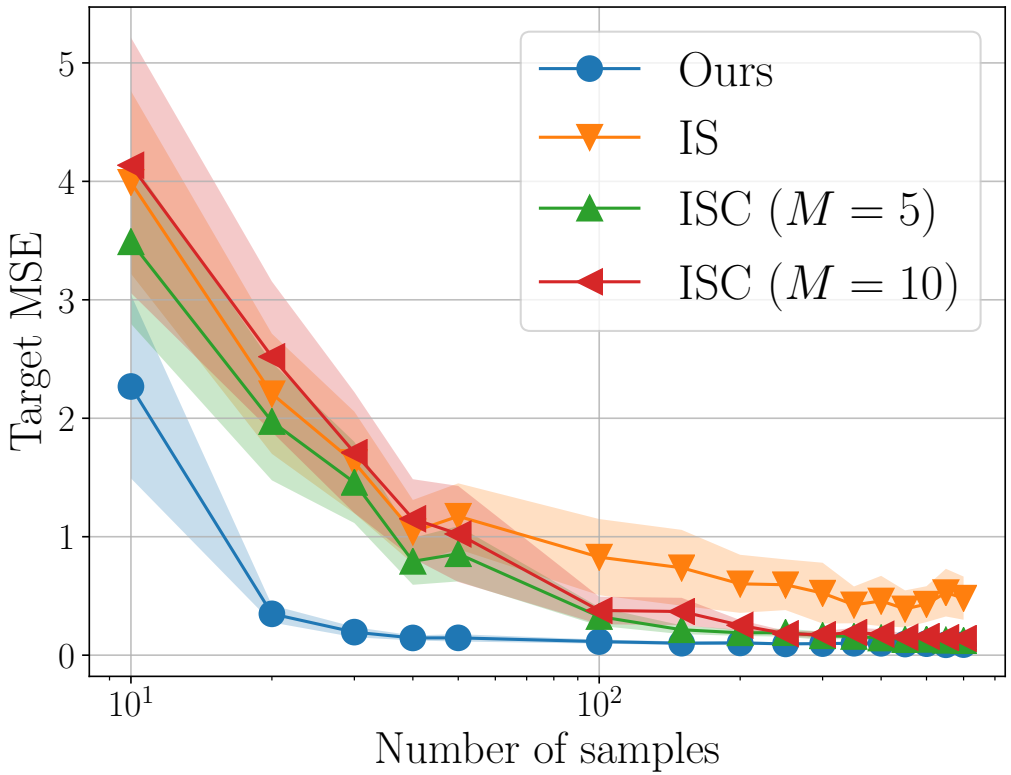

Figure: Target prediction error on synthetic domain adaptation experiment, comparing learned re-weighting (RCFR) and exact/clipped importance sampling weights (IS/ISC). Variance of IS hurts performance for small sample sizes.

"Real" data experiment - IHDP100

Table: Causal effect estimation on IHDP. CATE error RMSE($\hat{\tau}$), target prediction error $\hat{R}_\pi(f)$ and std errors. Lower is better.

	RMSE($\hat{\tau}$)	$\hat{R}_\pi(f)$
OLS	$2.3 \pm .11$	$1.1 \pm .05$
OLS-IPW	$2.4 \pm .11$	$1.2 \pm .05$
Random For.	$6.6 \pm .30$	$4.1 \pm .18$
Causal For.	$3.8 \pm .18$	$1.8 \pm .08$
BART	$2.3 \pm .10$	$1.7 \pm .07$
IPM-WNN	$1.2 \pm .12$	$.65 \pm .02$
CFRW	$.76 \pm .02$	$.46 \pm .01$
RCFR Oracle α, $w = 1$	$.81 \pm .07$	$.47 \pm .03$
RCFR Oracle α	$.65 \pm .04$	$.38 \pm .01$
RCFR Adapt. α	$.67 \pm .05$	$.37 \pm .01$

Conclusion

- The authors proposed a method to predict outcomes under shifts in design: changes in both policy and domain. This problem setting encompasses both domain adaptation and counterfactual prediction.
- Their framework uses representation learning, as well as a learned weighting function (from representation space), optimized in alternating fashion.
- They show the advantage of their method for domain adaptation, especially in the low-sample regime, even as compared with using closed-form importance sampling weights. They also show near-SOTA results on the standard IHDP100 benchmark.