A FRAMEWORK FOR GENERALIZATION AND TRANSPORTATION OF CAUSAL ESTIMATES UNDER COVARIATE SHIFT

APOORVA LAL, WENJING ZHENG, AND SIMON EJDEMYR

Randomized experiments are an excellent tool for estimating *internally valid* causal effects with the sample at hand, but their *external* validity is frequently questioned. While classical results on the estimation of Population Average Treatment Effects (PATE) implicitly assume random selection into experiments, this is typically far from true in many medical, social-scientific, and industry experiments. When the experimental sample is different from the target sample along observable or unobservable dimensions (termed *covariate shift* in the causal learning literature), experimental estimates may be of limited use for policy decisions. We cast this as a sample selection problem and propose methods to re-weight the doubly-robust scores from experimental subjects to estimate treatment effects in the overall sample (=: *generalization*) or in an alternate target sample (=: *transportation*). We implement these estimators in the open-source package `causalTransportR`[1] and illustrate its performance in a simulation study and discuss diagnostics to evaluate its performance.

Methods

We observe n iid copies of $(\mathbf{X}_i, S_i, S_i A_i, S_i Y_i)_{i=1}^n$, where covariates $\mathbf{X}_i \in \mathbb{R}^p$, treatment $A_i \in \mathcal{A} := \{0, \ldots, K\}$, outcome $Y_i \in \mathbb{R}$, and selection indicator $S_i \in \{0, 1\}$ is a function of pre-treatment variables and is not affected by treatment. In other words, we observe $(\mathbf{X}_i, A_i, Y_i)_{i=1}^{N_1}$ for observations with $S_i = 1$ (henceforth the *study* sample $\mathcal{S}_1$), and only $(\mathbf{X}_i)_{i=N_1+1}^{N}$ for observations with $S_i = 0$ (henceforth the *external* sample $\mathcal{S}_0$). The *overall* sample is $\mathcal{S} := \mathcal{S}_1 \cup \mathcal{S}_0$.

Estimands. We write counterfactual means as $\phi = \mathbb{E}\left[Y^{a,S=1}\right]$ for generalizability and $\mathbb{E}\left[Y^a | S = 0\right]$ for transportability, and contrasts between such counterfactual means under any two treatment levels a, a' represent the average treatment effects (ATE). 'Standard' estimation of effects in the study sample under unconfoundedness is a well-studied and largely resolved problem (see [10] for a review). We study the generalization and transportation problems in the present paper. To this end, we make the following assumptions:

(1) Consistency / SUTVA : $Y_i = 1_{A_i=a} Y_i^a$
(2) Ignorability of Treatment: $Y^0, \ldots, Y^a \perp\!\!\!\perp A | X = x, S = 1$
(3) Overlap
    (a) Treatment overlap: $0 < \mathbf{Pr}\left(A = a | \mathbf{X} = \mathbf{x}, S = 1\right) < 1$

E-mail addresses: apoorval@stanford.edu, wzheng@netflix.com, sejdemyr@netflix.com.
Date: October 22, 2022.
Key words and phrases. experimentation, generalization, transportation, bridging.
[1]Available at https://github.com/Netflix-Skunkworks/causalTransportR

   (b) Selection overlap: $0 < \mathbf{Pr}(S = 1|\mathbf{X} = \mathbf{x}) < 1$
(4) Selection
   (a) $Y^0, \ldots, Y^a \perp\!\!\!\perp S|\mathbf{X} = \mathbf{x}$ Ignorability of Selection.
   (b) $\mathbb{E}[Y|A, \mathbf{X}, S = 1] = \mathbb{E}[Y|A, \mathbf{X}, S = 0]$. The outcome model is stable across S strata.

Under assumptions 1,2,3 and 4a, causal quantities of interest in the overall sample are identified [2, 6], while under 1,2,3, and 4b, causal quantities of interest are identified in the target sample [5]. While prior work focused on binary treatments, we establish identification and estimation for counterfactual means and causal contrasts for multiple discrete treatments, which is the norm at Netflix and other industry settings.

Estimators. Our preferred estimators are efficient influence function (EIF) based that take the form of sample averages $\widehat{\psi} = \frac{1}{n}\sum_{i=1}^{n}\varphi(W_i)$ where $W_i = (A_i, \mathbf{X}_i, Y_i, S_i)$. The influence function obeys $n^{1/2}(\widehat{\psi} - \psi) = n^{-1/2}\sum_{i=1}^{n}\varphi(W_i) + o_p(1)$. This form characterizes Regular and Asymptotically Linear (RAL) estimators and allows us to construct valid confidence intervals using the sample variance of the influence function. These estimators rely on the estimation of three nuisance functions: (1) Outcome model $\mu^a(\mathbf{x}) = \mathbb{E}[Y|A = a, \mathbf{X} = \mathbf{x}]$, (2) Treatment Propensity score $\pi^a(\mathbf{x}) = \mathbf{Pr}(A = a|\mathbf{X} = \mathbf{x}, S = 1)$, and (3) Selection propensity score $\rho(x) = \mathbf{Pr}(S = 1|\mathbf{X} = \mathbf{x})$. Their sample analogues $\widehat{\alpha}$ are fit using machine learning estimators with cross-fitting and is implemented in the package with regularized regressions and generalized random forests.

The estimators under consideration take on one of three forms outlined in table 1. **Outcome Modeling (OM)** is a pure transfer-learning approach that involves fitting conditional response surfaces $\mathbb{E}[Y|A = a, S = 1]$ over the observations with nonmissing Y and extrapolating these over the relevant samples. **Inverse Selection Weighting (ISW)** involves modeling the selection probability into the source sample with covariates, and reweighting observations to mimic the target samples. **Augmented ISW (AISW)** combines the ISW and OM approaches by augmenting the outcome model with an weighted average of residuals $(Y - \mu(\cdot))$ and possesses double-robustness properties (from outcome of both propensity models) analogous to the classical Augmented IPW estimator [14]. Both ISW and AISW can be stabilized using a Hajek normalization term equal to the sum of weights in each treatment level a.

If only summary statistics are available for the target sample, AISW is infeasible since it requires individual level covariates for all observations to construct weights. In such cases, a 'calibration' approach based on solving for balancing weights that ensure balance between two population is feasible and has appealing properties in finite samples , and is also implemented using entropy loss [9] in the package.

TABLE 1. Estimators. Difference between marginal means $\widehat{\psi}_a - \widehat{\psi}_{a'}$ yields causal contrasts $\tau(a, a')$. Standard errors are computed as $\sqrt{\hat{\sigma}^2/n}$ where $\widehat{\sigma^2}$ is the sample variance of the influence function of interest (marginal mean or causal contrast) for AISW, or via the nonparametric or bayesian bootstrap for other estimators

	Generalization $\sum_x \mathbb{E}\left[Y \mid A = a, S = 1, \mathbf{X}\right] P(\mathbf{X})$	**Transportation** $\sum_x \mathbb{E}\left[Y \mid A = a, S = 1, \mathbf{X}\right] P(\mathbf{X} \mid S = 0)$
OM	$\frac{1}{n}\sum_i \widehat{\mu}^a(\mathbf{X}_i)$	$\frac{1}{\lvert\mathcal{S}_0\rvert}\sum_i (1 - S_i)\widehat{\mu}^a(\mathbf{X}_i)$
ISW	$\frac{1}{n}\sum_i \frac{S_i}{\widehat{\rho}(\mathbf{X}_i)} \frac{1_{A=a}}{\widehat{\pi}^a(\mathbf{X}_i)} Y_i$	$\frac{1}{n}\sum_i \frac{1}{\widehat{\mathbb{E}}[S_i=0]} \frac{S_i(1-\widehat{\rho}(\mathbf{X}_i))}{\widehat{\rho}(\mathbf{X}_i)} \frac{1_{A=a}}{\widehat{\pi}^a(\mathbf{X}_i)} Y_i$
AISW	$\frac{1}{n}\sum_i \widehat{\mu}^a(\mathbf{X}_i) + \frac{S_i}{\widehat{\rho}(\mathbf{X}_i)} \frac{1_{A=a}}{\widehat{\pi}^a(\mathbf{X}_i)} (Y_i - \widehat{\mu}^a(\mathbf{X}_i))$	$\frac{1}{n}\sum_i \frac{1}{\widehat{\mathbb{E}}[S_i=0]} \left((1 - S_i)\widehat{\mu}^a(\mathbf{X}_i) + \frac{S_i(1-\widehat{\rho}(\mathbf{X}_i))}{\widehat{\rho}(\mathbf{X}_i)} \frac{1_{A=a}}{\widehat{\pi}^a(\mathbf{X}_i)} (Y_i - \widehat{\mu}^a(\mathbf{X}_i)) \right)$

SIMULATION STUDY

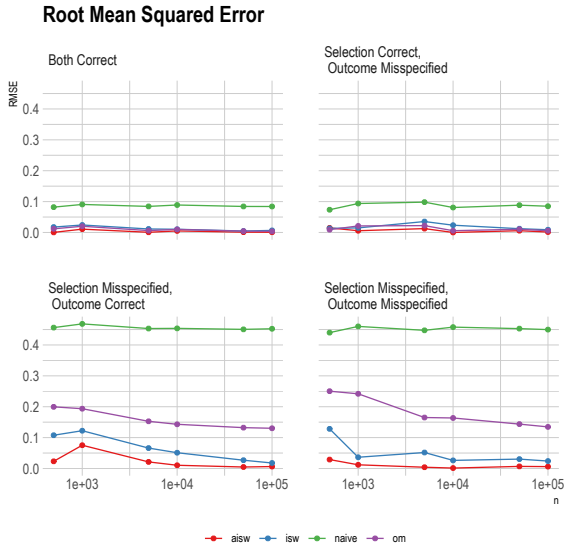
We study the generalization estimators' performance in a simulation study, where we simulate four scenarios where covariates $x_1, \ldots, x_{10} \sim \mathrm{U}[-1, 1]$ and treatment is randomly assigned with probability 0.5, and the true (selection / outcome) models are (linear / nonlinear), and nuisance functions are estimated using regularized linear regressions with λ set to minimise CV-MSE. The selection model dictates how the study sample is selected from the target population, and whenever this is a function of covariates, the experimental estimate of the average treatment effect (SATE) is biased for the treatment effect in the target population (PATE).

We report the performance of the above estimators in figure 1, which displays the RMSE, Bias, Coverage rate, and runtime across 500 replications. We find that reweighting estimators (red, blue, and purple) consistently outperform 'naive' SATE estimates (green). Consistent with the analogous results in the unconfoundedness literature, we find that among the generalization estimators, the augmented inverse selection weighting estimator (red) performs best along MSE, bias, and variance dimensions.

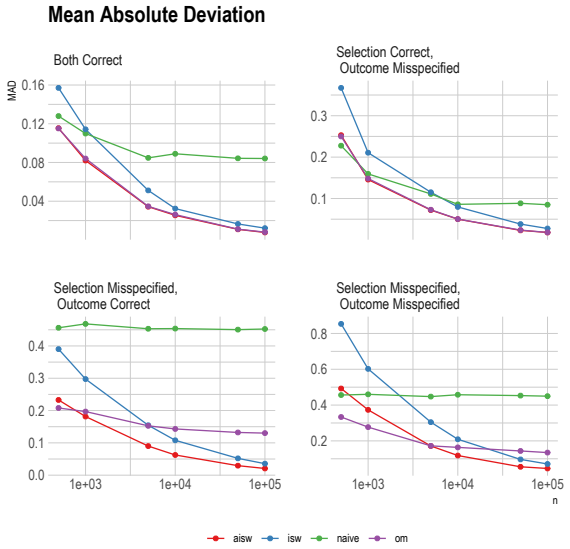
DISCUSSION

In this work, we provide a concise multi-treatment framework for causal generalization and transportation, and provide a performant computational implementation for it. In current practice, practitioners often informally perform 'naive' extrapolation of the Sample Average Treatment Effect (SATE) to the Target Average Treatment Effect (TATE), which may result in erroneous conclusions arising from three distinct sources of bias generated by differences between study and external samples : (1) unequal distribution of effect modifier covariates in the two samples, (2) Lack of covariate overlap between the two

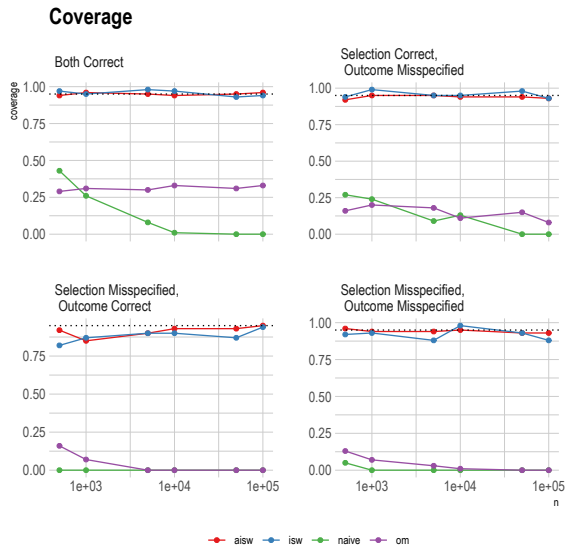
FIGURE 1. Simulation Results



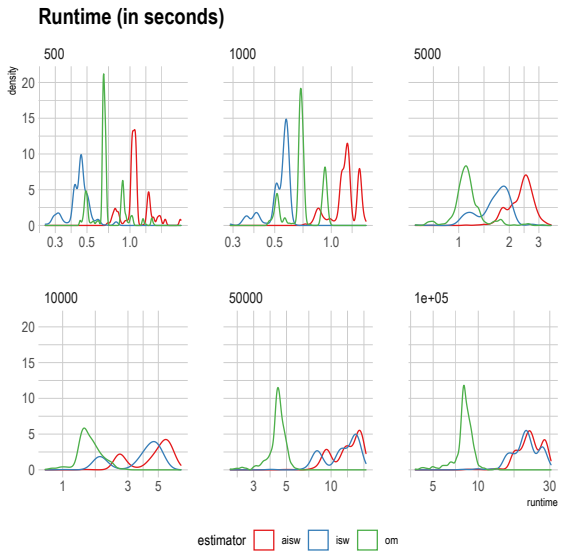
(A) RMSE



(B) Bias



(C) 95% CI coverage



(D) Runtime

samples, and (3) differences between the effect modification functions between the two samples[2].

[2]We provide a fuller derivation of these in appendix 1

Naive extrapolation can be improved upon by using the framework presented in the present paper, which shows that causal generalization and transportation involves (1) estimating strata-specific Conditional Average Treatment Effects (CATEs) $\tau_s(\mathbf{X})$, (2) asserting outcome model stability wherein the heterogeneity function $\tau(\mathbf{X})$ is stable across the study and external sample, and (3) reweighting CATEs to match the covariate distribution $p(\mathbf{X})$ in the target population. The feasibility of each of these steps may be problem-specific, and practitioners are advised to carefully consider potential problems in each step in their particular application.

While an enormous literature has emerged on CATE estimation (see [11] for a review), evaluating the quality of the resultant estimates remains a challenging problem. Omnibus tests for systematic treatment effect heterogeneity [7](implemented as `dfmTest` in the package) and [3] are strongly recommended prior to the use of generalization estimators. CATE estimation for settings with small treatment effects as is typical in industry settings is a growing area of research [1] and progress on this problem can readily be integrated to improve upon effect transportation.

Outcome model stability across the study and external samples is inherently untestable, and must be justified from substantive knowledge of the study and target population, as well as the nature of the experimental manipulation. For example, a video compression treatment might have systematic and stable treatment effects across subpopulations, while a recommendation algorithm may not because of preference heterogeneity that isn't adequately captured by covariates. Sensitivity analyses in the vein of [4, 8, 12] that assess the magnitude of violations of outcome model stability that overturn transportation conclusions are a fruitful avenue for future research.

Finally, the effectiveness of selection weights in reducing the imbalance between the study and external populations can be evaluated using the suite of tools developed for propensity score weights. Plotting standardized mean differences between the source and target samples is a reasonable first check for whether the weights are effective at reducing imbalance, and is implemented as the plot method for the model object in the package. These figures are also also intended to help choose between 'indirect' balancing via propensity score modelling versus calibration approaches that target balance directly.

References

[1] Susan Athey et al. *Semiparametric Estimation of Treatment Effects in Randomized Experiments*. Tech. rep. National Bureau of Economic Research, 2021.

[2] Michela Bia, Martin Huber, and Lukáš Lafférs. "Double machine learning for sample selection models" (Nov. 2020). arXiv: 2012.00745 [econ.EM]. URL: http://arxiv.org/abs/2012.00745.

[3] Victor Chernozhukov et al. *Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments*. 2020. arXiv: 1712.04802 [stat.ML].

[4] Victor Chernozhukov et al. *Long story short: Omitted variable bias in causal machine learning*. Tech. rep. National Bureau of Economic Research, 2022.

[5] Issa J Dahabreh et al. "Extending inferences from a randomized trial to a new target population". en. *Statistics in medicine* 39.14 (June 2020), pp. 1999–2014.
[6] Issa J Dahabreh et al. "Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals". en. *Biometrics* 75.2 (June 2019), pp. 685–694.
[7] Peng Ding, Avi Feller, and Luke Miratrix. "Decomposing Treatment Effect Variation". *Journal of the American Statistical Association* 114.525 (Jan. 2019), pp. 304–317. URL: https://doi.org/10.1080/01621459.2017.1407322.
[8] Jacob Dorn and Kevin Guo. "Sharp sensitivity analysis for inverse propensity weighting via quantile balancing". *Journal of the American Statistical Association* just-accepted (2022), pp. 1–28.
[9] Jens Hainmueller. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies". *Political analysis: an annual publication of the Methodology Section of the American Political Science Association* 20.1 (2012), pp. 25–46.
[10] Edward H Kennedy. "Semiparametric doubly robust targeted double machine learning: a review" (Mar. 2022). arXiv: 2203.06469 [stat.ME]. URL: http://arxiv.org/abs/2203.06469.
[11] Michael C Knaus, Michael Lechner, and Anthony Strittmatter. "Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence". en. *The econometrics journal* 24.1 (June 2020), pp. 134–161.
[12] Xinkun Nie, Guido Imbens, and Stefan Wager. "Covariate Balancing Sensitivity Analysis for Extrapolating Randomized Trials across Locations" (Dec. 2021). arXiv: 2112.04723 [econ.EM]. URL: http://arxiv.org/abs/2112.04723.
[13] My Phan et al. "Designing Transportable Experiments Under S-admissability". *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, 13–15 Apr 2021, pp. 2539–2547.
[14] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. "Estimation of regression coefficients when some regressors are not always observed". *Journal of the American statistical Association* 89.427 (1994), pp. 846–866.

Part 1. Appendix

Bias decomposition

To simplify notation, lets assume covariate $\mathbf{X}_i \in \mathcal{X}$ is discrete, and follows distributions $p_s(\mathbf{X})$ and $p_t(\mathbf{X})$ in the study and target samples respectively. We denote the Conditional Average Treatment Effect (CATE) as $\tau_k(\mathbf{x}), k \in \{s, t\}$. We assume CATE stability across samples $\tau_s(\mathbf{x}) = \tau_t(\mathbf{x}) = \tau(\mathbf{x})$, and discuss implications of relaxing this assumption next.

The gap between the Target Average Treatment Effect (TATE) and Sample Average Treatment Effect (SATE) can be decomposed as follows

$$\begin{aligned}
\text{TATE - SATE} &= \sum_{x \in \mathcal{X}_t} p_t(\mathbf{x})\tau_t(\mathbf{x}) - p_s(\mathbf{x})\tau_s(\mathbf{x}) \\
&= \sum_{x \in \mathcal{X}_t} \left(p_t(\mathbf{x}) - p_s(\mathbf{x})\right)\tau(\mathbf{x}) && \text{by } \tau_s(\cdot) = \tau_t(\cdot) = \tau(\cdot) \\
&= \sum_{x \in \mathcal{X}_t} p_s(\mathbf{x}) \left(\frac{p_t(\mathbf{x})}{p_s(\mathbf{x})} - 1\right)\tau(\mathbf{x})
\end{aligned}$$

From the above, we can see three distinct sources of bias:

(1) When overlap holds, bias contributions come from strata where the following three conditions are true
   (a) $\tau(\mathbf{x}) \neq 0$: Non-zero treatment effects
   (b) $p_s(\mathbf{x}) > 0$: Nonzero support in study population
   (c) $p_t(\mathbf{x}) \neq p_s(\mathbf{x})$: Distribution of covariate $\mathbf{x}$ is different across study and target population
(2) When overlap is violated such that $p_t(\mathbf{x}) > 0$ and $p_s(\mathbf{x})$: there exist strata in the target population that are unrepresented in the study, the SATE is biased and the bias is increasing in the size of $p_t(\mathbf{x})$ (fraction of target sample unrepresented in study population) and $\tau(\mathbf{x})$
(3) If the CATE functions $\tau_s(\mathbf{x}) \neq \tau_s(\mathbf{x})$ (i.e. effect modification is different between the study and target sample).

Our reweighting approach addresses (1) using a post-stratification weights for discrete covariates and its analogue selection score for continuous covariates. (2) and (3) produce bias that is impossible to resolve without additional data collection (for 2, for example through the use of S-admissable designs [13]) or prior knowledge of CATE functions (for 3).