Generalization Bounds for Causal Regression:
Insights, Guarantees and Sensitivity Analysis

Daniel Csillag, Claudio J. Struchiner and Guilherme T. Goedert
School of Applied Mathematics of Getulio Vargas Foundation, Rio de Janeiro, Brazil

Contact info:
@dccsillag
dccsillag@gmail.com
https://dccsillag.xyz/

Introduction

At the core of causal ML is reasoning about potential outcomes Y^a, which correspond to *what would happen* to the outcome Y if we were to intervene and make $T = a$. This is subtly different from simply predicting which observed outcome Y is most likely — correlation is not causation!

Many methods have been proposed for causal ML, but generally with a rather shaky theoretical background and with heavy assumptions. So which ones actually, provably work? And which ones are robust to violations of their assumptions? In this work, we sought to develop a new backing theory for such methods based on the formalism of bounds on the generalization error.

Actually important: In causality we can't just resort to benchmarking because we fundamentally cannot observe the ground truths. We *need* theory!

Our goal: introduce generalization bounds for causal regression:

$$\text{test causal loss} \leq \text{train observable loss} + \text{fitting complexity} \\ + \underline{\text{causal penalty}} + O\big(n^{-1/2}\big).$$

We consider two key tasks of causal ML:

(i) **potential outcome regression**, in which we seek to predict the potential outcomes Y^a given the covariates X; and
(ii) **individual treatment effect estimation**, in which we seek to predict the treatment effects $Y^1 - Y^0$ given the covariates X.

Assumptions

- **I.I.D. data** *(can be relaxed)*
  (a.k.a. no-interference in the causal inference literature.)
- **Consistency** *(can be relaxed)*
  Conditional on $T = a$, it holds that $Y^a = Y$, for all values of a.
- **Ignorability** w.r.t. observed and unobserved covariates
  $$(Y^1, Y^0) \perp\!\!\!\perp T \mid X, U.$$

This is all we need for bounds for estimation of potential outcomes. For bounds for the estimation of treatment effects we introduce one additional light assumption on the loss function.

From Observable Losses to Causal Losses

First we bound the loss in expectation:

Theorem 2.3. For any loss function, reweighting function $w(X)$ with $\mathbb{E}[w(X)] = 1$ and any $\lambda > 0$,

$$\underbrace{\mathbb{E}[\text{Loss}]}_{\text{unobservable...}} \leq \underbrace{\mathbb{E}[w(X) \cdot \text{Loss} \mid T = a]}_{\text{observable!}} + \lambda \Delta_{T=a} + \sigma^2_{T=a}/4\lambda$$

where

$$\Delta_{T=a} = \mathbb{E}\left[\left(w(X)\frac{\mathbb{P}[T = a \mid X, U]}{\mathbb{P}[T = a]} - 1\right)^2\right] \text{ and } \sigma^2_{T=a} = \text{Var}[\text{Loss}].$$

Proof sketch. By using a lemma we introduce in the paper, we can tightly bound the gap between expectations of any two distributions by their χ^2 divergence. Applying it to the distributions of P_{Loss} and $P_{\text{Loss} \mid T=a}$ gets us to the desired inequality except for having $\Delta_{T=a} = \chi^2\big(P_{Y,X,U|T=a} \parallel P_{Y^a,X,U}\big)$.

Since the χ^2 divergence is an f-divergence, it is a function of the density ratio of its inputs. And indeed, in the causal setting, by consistency and ignorability, we have that $\mathrm{d}P_{Y,X,U|T=a}/\,\mathrm{d}P_{Y^a,X,U} = w(X)\mathbb{P}[T = a|X, U]/\mathbb{P}[T = a]$. Plugging this into the definition of the χ^2 divergence concludes the proof. ■

Unfortunately, $\Delta_{T=a}$ is unknown in practice due to $\mathbb{P}[T = a|X, U]$. But, quite remarkably, we can bound it in a way we can estimate *with no knowledge of* U:

Theorem 2.4. For $\Delta_{T=a}$ as in Theorem 1,

$$\Delta_{T=a} \leq \frac{2}{\mathbb{P}[T = a]^2} \cdot \Big(\mathbb{E}\Big[w^2(X) \cdot (\nu(X) - \mathbb{1}[T = a])^2\Big] \\ + \mathbb{E}\Big[(w(X)\mathbb{1}[T = a] - \mathbb{P}[T = a])^2\Big]\Big).$$

Proof sketch. By adding and subtracting $\mathbb{1}[T = a]/\mathbb{P}[T = a]$ and using a relaxed triangular inequality, we get the desired inequality except for $\nu(X)$ being replaced by $\mathbb{P}[T = a|X, U]$.

The proof is concluded by noting that the first expectation is a reweighted Brier loss of $\mathbb{P}[T = a|X, U]$ w.r.t. $\mathbb{1}[T = a]$, which is optimized precisely for $\mathbb{P}[T = a|X, U]$. Therefore, substituting it for $\nu(X)$ keeps the bound valid. ■

Generalization Bounds and More

Generalization Bounds for Prediction of Potential Outcomes: We can combine Theorems 2.3 and 2.4 with generalization bounds for the non-causal case in order to achieve generalization bounds for causal regression. We give an example with Rademacher-based bounds, but the same idea applies to other frameworks (e.g., PAC-Bayes, VC, stability, etc.).

Corollary 2.5. For any loss function bounded in $[0, M]$ and reweighting w as in Theorem 2.3 bounded in $[0, w_{\max}]$, for any $\lambda > 0$, with high probability, for any model $h \in \mathcal{H}$,

$$\overbrace{\mathbb{E}[\text{Loss}(h)]}^{\text{test causal loss}} \leq \overbrace{n^{-1}_{T=a} \sum_{T_i=a} w(X_i)\text{Loss}_i(h)}^{\text{train observable loss}} + \overbrace{2\mathfrak{R}(\mathcal{H}) + 2\mathfrak{R}(\mathcal{H}_\nu)}^{\text{fitting complexity}} \\ + \underbrace{\lambda\hat{\Delta}_{T=a} + M^2/16\lambda}_{\text{causal penalty}} + O\Big(Mw_{\max} \cdot n^{-1/2}_{T=a}\Big).$$

Generalization Bounds for Prediction of Treatment Effects: By introducing an additional light assumption on the structure of the loss function, we can separate the loss of treatment effect predictors into losses of individual potential outcome regressions. We can then leverage the bounds we've developed for these individual regressions along with an union bound to get generalization bounds for the treatment effect estimation problem. See Section 2.2 of the paper!

Prediction of Treatment Effects Beyond the MSE: One remarkable aspect of our bounds is that they are loss-agnostic: i.e., they hold not only for the MSE loss, but also for 0-1 loss (for classification), MAE (for robust regression) and, most remarkably, for the quantile loss (for quantile regression). This is quite notable, since it shows that it *is* possible to estimate conditional quantiles of treatment effects, contrary to common belief. See Section 2.3 of the paper!

Experiments on Semi-Synthetic Data: We conduct experiments on datasets of varying complexity, showcasing the remarkable tightness of our bounds. Not only are they tight (even when there is hidden confounding!), but they are also *orders of magnitude* tighter than the closest matching result previously available in the literature. See Section 3.1 of the paper!

Experiments on Real Data: We futher demonstrate the practical utility of our bounds by showcasing their effectiveness on a model selection task on tricky real data full of hidden confounding. See Section 3.2 of the paper!