Targeted Learning

Yuri Fonseca

April 19, 2023

1 Introduction

In this lecture note we discuss the basics of targeted maximum likelihood estimator (TMLE). The note is based on Van der Laan et al. (2011); Kennedy (2022) and the Machine-Learning and Causality reading group[1] at Columbia University. TMLE is a *nonparametric* method that a researcher can use to answer *queries* that take as an input *observational data* from an *unknown distribution*. An important application is in causal inference where, under the additional assumptions of a Structural Causal Model *causal queries* from observational data can be answered. This is specially important in settings where experimental studies are expensive/not expensive and *structural* assumptions on the data generating process are likely to be wrong.

2 Notation and Basic Definitions

We have iid copies of the random variable denoted by $D = \{X, A, Y\} \sim P_0$. $\{D_i\}_{i=1}^n$ is our data. The true distribution P_0 is unknown. We assume a statistical model $\mathcal{M}$. We also assume that $P_0 \in \mathcal{M}$. One could also impose a structure in the statistical model by defining $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ where θ itself is infinite dimensional.

We are interested in evaluating a *target parameter* (a functional) ψ under the true distribution. I.e., for $\psi : \mathcal{M} \to \mathbb{R}^d$ we want to compute $\psi_0 \triangleq \psi(P_0)$. For concreteness, we focus in the case where treatments are binary, and we want to compute

$$\psi_0 = \mathbb{E}_{X,0}[\mathbb{E}_0[Y|A=1,X] - \mathbb{E}_0[Y|A=0,X]],$$

where the expectation is under the true distribution P_0. We will see examples in which Y is binary, and continuous and bounded. Recall our *nontestable* causal assumption (ignorability). We have that

$$\begin{aligned}
\mathbb{E}_0[Y(1) - Y(0)] &\overset{(a)}{=} \mathbb{E}_{X,0}[\mathbb{E}_0[Y(1) - Y(0)|X]] \\
&= \mathbb{E}_0[\mathbb{E}_0[Y(1)|X] - \mathbb{E}_0[Y(0)|X]] \\
&\overset{(b)}{=} \mathbb{E}_0[\mathbb{E}_0[Y(1)|X, A=1] - \mathbb{E}_0[Y(0)|X, A=0]] \\
&= \mathbb{E}_0[\mathbb{E}_0[Y|X, A=1] - \mathbb{E}_0[Y|X, A=0]] \\
&= \psi_0,
\end{aligned}$$

[1] Special thanks to Judy Gan and Tiffany Cai.

where (a) follows from the tower property and (b) from ignorability. Hence, under causal assumptions, TMLE help us to actually answer the *causal query:* Average Treatment Effect (ATE). As before, we use the notation

$$\mu(A=1,X)=\mathbb{E}_0[Y|A=1,X], \quad \mu(A=0,X)=\mathbb{E}_0[Y|A=0,X],$$

we also denote by $P_{X,0}$ the marginal distribution of X under P_0. Note that is enough for us to know only

$$Q_0=(\mu,P_{X,0}),$$

in order to compute $\psi(P_0)$. Since we focus only on learning Q_0 instead of the whole P_0, we denote it as our *target*, and TMLE is a methodological procedure that will allow us to move from an initial estimator $\hat{Q}_n$[2] of Q_0 to a **targeted** one that is closer to Q_0 that we denote by $Q_n^\star$.

3 Debiasing

Suppose our target parameter (functional) ψ satisfy the following expansion for two distributions $\bar{P}$ and P:

$$\psi(\bar{P})-\psi(P)=\int \nabla\psi(\bar{P})d(\bar{P}-P)+R_2(\bar{P},P)$$

which is referred to as *Von Mises expansion* and $\nabla\psi(P)$ is a square integrable function with zero mean and $R_2(\bar{P},P)$ is a second order term. You can think of it as being a Taylor expansion of ψ and $\nabla\psi$ being the derivative of ψ, which is commonly referred to as the *Efficient Influence Function/Curve* (IF/IC) of ψ or pathwise derivative of ψ. Other common notations for the IF are ϕ or $\dot{\psi}$.

Example 1. *Let $\psi(P)=\mathbb{E}[Y|A=1,X]$. Then, the Influence Function of ψ is given by*

$$\nabla\psi(P)(Y,A,X)=\frac{I(A=1)}{e(A=1|X)}(Y-\mu(1,X))+\mu(1,X)-\psi(P).$$

This expansion is very powerful, and allows us to characterize the bias of a plug-in estimator based on some initial estimator $\hat{P}_n$. We have that

$$\begin{aligned}\psi(\hat{P}_n)-\psi(P_0)&=\int\nabla\psi(\hat{P}_n)d(\hat{P}_n-P_0)+R_2(\hat{P}_n,P_0)\\&=\int\nabla\psi(\hat{P}_n)d\hat{P}_n-\int\nabla\psi(\hat{P}_n)dP_0+R_2(\hat{P}_n,P_0)\\&=-\int\nabla\psi(\hat{P}_n)dP_0+R_2(\hat{P}_n,P_0),\end{aligned}$$

where the last inequality follows from the fact that $\nabla\psi$ is zero mean. Suppose now that we have access to another sample independent of $\hat{P}_n$ that we denote by P_{n_2}. Since the RHS is just an expectation, we can "pretend" that

$$\psi(\hat{P}_n)+\int\nabla\psi(\hat{P}_n)dP_0\approx\psi(\hat{P}_n)+\frac{1}{n_2}\sum_i\nabla\psi(\hat{P}_n)(D_i),$$

[2]We use the subscript n to show the dependence of the sample size.

which is an unbiased plug-in estimator, and $\frac{1}{n}\sum_i \nabla\psi(\hat{P}_n)(D_i)$ is the unbiasing term.

3.1 AIPW

Now we show how to debias a naive plug-in estimator using the *Von Mises* expansion. As we will see shortly, this leads to the AIPW (doubly-robust and assymptotically optimal) estimator. We start by solving

$$\mu_{\hat{P}_n} = \arg\min_{\mu} \frac{1}{n}\sum_i (Y_i - \mu(A_i, X_i))^2,$$

and the naive plug-in estimator is simply

$$\psi(\hat{P}_n) = \frac{1}{n}\sum_i \mu(1, X_i) - \mu(0, X_i),$$

which is likely to be biased. Next, we compute on a different sample n_2, $\frac{1}{n_2}\sum_i \nabla\psi(\hat{P}_n)(D_i)$, and summing both terms we get:

$$\begin{aligned}\psi(\hat{P}_n) + \frac{1}{n_2}\sum_i \nabla\psi(\hat{P}_n)(D_i) = \psi(\hat{P}_n) - \frac{1}{n_2}\sum_i \psi(\hat{P}_n) + \frac{1}{n_2}\sum_i (\hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)) \\ + \frac{1}{n_2}\sum_i \left(\frac{I(A_i = 1)}{e_{\hat{P}_n}(A_i = 1|X_i)} - \frac{I(A_i = 0)}{e_{\hat{P}_n}(A_i = 0|X_i)}\right)(Y_i - \hat{\mu}(A_i, X_i))\end{aligned}$$

which is precisely AIPW when we use a two split procedure. Note that it can be interpreted as a standard plug-in estimator plus a first order correction term.

3.2 TMLE

At the heart of TMLE, just like in the debiasing procedure described before, is the efficent influence function. Recall that the IF for the ATE under binary treatment is given by

$$\nabla\psi(Q_0)(Y, A, X) = \left(\frac{I(A = 1)}{e_0(A = 1|X)} - \frac{I(A = 0)}{e_0(A = 0|X)}\right)(Y - \mu(A, X)) + \mu(1, X) - \mu(0, X) - \psi(Q_0).$$

Let $\hat{Q}_n = (\hat{\mu}, \hat{Q}_X)$ be our initial estimator for our region of interest of P_0. Define a loss function L such that $\mathbb{E}_0[L(Q)]$ is minimized at Q_0 (the relevant part of the true distribution). As an example, one could take $L(P) = -\log P$ (in this case, $\mathbb{E}_0[L(P)]$ is minimized at P_0). Next, Recall that for any $P \in \mathcal{M}$, we can define a parametric submodel $\{P(\epsilon) : \epsilon\}$ so that $P(0) = P$. Therefore, we can take our initial estimator for relevant part of P_0 and define $\{\hat{Q}_n(\epsilon)\}$ with $\hat{Q}_n(0) = \hat{Q}_n$. Until now there is nothing really important going on. Here is the smart part: We also want that our parametric submodel have the property that $\frac{d}{d\epsilon}L(\hat{Q}_n(\epsilon))$ (the *score*) evaluated at $\epsilon = 0$ matches the influence function evaluated at $\hat{Q}_n$.

Why is this powerful? Because then we can solve the optimization problem $\epsilon^\star = \arg\min_\epsilon L(\hat{Q}_n(\epsilon))$ and ensure that for $\hat{Q}_n(\epsilon^\star)$, $\mathbb{E}_n[\nabla\psi(Q_0)] = 0$, i.e., in the empirical measure, $\hat{Q}_n(\epsilon^\star)$ is already debiased. Next, we give two examples in how to define a parametric submodel applies the procedure described above and allow us to debias our

estimator for the ATE efficiently. It will be useful to define the notation

$$H^\star(A,X) = \left(\frac{I(A=1)}{\hat{e}(A=1|X)} - \frac{I(A=0)}{\hat{e}(A=0|X)}\right).$$

The function $H^\star$ is sometimes refereed to as the clever covariate.

On the P_0 factorization. We want to build parametric submodels such that their scores spans the efficient IF. We can write

$$\nabla\psi(Q_0)(Y,A,X) = \left(\frac{I(A=1)}{e_0(A=1|X)}\right)(Y-\mu(1,X)) + \mu(1,X) - \psi(Q_0) = D_Y^\star + D_W^\star.$$

for $D_Y^\star = \left(\frac{I(A=1)}{e_0(A=1|X)}\right)(Y-\mu(1,X))$ and $D_X^\star = \mu(1,X) - \psi(Q_0)$. However, in the discussion above we presented how to construct a submodel to deal with $D_Y^\star$ only. Why we do not worry about updating $D_X^\star$ also?

Note that for $D_X^\star$, one could define the parametric submodel given by:

$$dP(\epsilon) = dP(1+\epsilon f(X)),$$

for $f(X) = \mu(1,X) - \psi(Q_0)$. This is a mean zero function, moreover, for $L(P(\epsilon))$ given by the negative log likelihood function, we get that $\frac{d}{d\epsilon}\log dP(\epsilon) = \frac{\mu(1,X)-\psi(Q_0)}{1+\epsilon(\mu(1,X)-\psi(Q_0))}$, which matches $D_X^\star$ when evaluated at $\epsilon = 0$, satisfying all the requirements for the TMLE procedure. Moreover, when optimizing $L(P(\epsilon))$, with respect to ϵ, we get from the FOC that $\epsilon = 0$, which is due to the fact that the empirical measure already minimizes the negative log likelihood on the observed data.

Example 2. *Suppose Y continuous, bounded and $A \in \{0,1\}$ We define $L(Q)(O) \triangleq (Y-\mu(A,X))^2$. Next we define our parametric submodel. Since $\hat{Q}_X$ is simply the nonparametric estimator for the marginal distribution of X, it is already assymptotically optimal. We would like to "debias" $\hat{\mu}$. We define the parametric submodel for μ as*

$$\mu_n(\epsilon)(A,X) = \hat{\mu}(A,X) + \epsilon H^\star(A,X).$$

Then,

$$\begin{aligned}\frac{d}{d\epsilon}L(\mu_n(\epsilon^\star)) &= -2(Y-\hat{\mu}(A,X) - \epsilon^\star H^\star(A,X))H^\star(A,X) \\ &= \left(\frac{I(A=1)}{\hat{e}(A=1|X)} - \frac{I(A=0)}{\hat{e}(A=0|X)}\right)(Y-\hat{\mu}^\star(A,X)) \\ &= 0,\end{aligned}$$

where the last equality follows from defining $\mu^\star(A,X) \triangleq \hat{\mu}(A,X) + \epsilon^\star H^\star(A,X)$.

Therefore, we can debias a initial estimator by solving an one-dimensional optimalization problem to find $\epsilon^\star$, and update our initial estimator. Note that the FOC for the optimazation problem that we solve for ϵ, already implies that $\mu^\star$ is unbiased since it "solves" for the relevant part of the IF.

Next we provide an example in which the outcomes are binary and we use a different loss function to define our submodel.

Example 3. *Suppose $Y \in \{0,1\}$ and $A \in \{0,1\}$. We define $L(Q)(O) \triangleq -\log \mu(A,X)^Y(1-\mu(A,X))^{1-Y}$, i.e., we take the (minus) of the log-likelihood function for the negative binomial. Next we define our parametric submodel. Since $\hat{Q}_X$ is simply the nonparametric estimator for the marginal distribution of X, it is already assymptotically optimal. We would like to "debias" $\hat{\mu}$. We define our submodel as*

$$logit\, \mu_n(\epsilon)(A,X) = logit\, \hat{\mu}(A,X) + \epsilon H^\star(A,X),$$

*Then, since the score of a logistic regression parameter is the error times the covariate (**add reference**), we have that*

$$\frac{d}{d\epsilon} L(\mu_n(\epsilon)) = (Y - \hat{\mu}(A,X))H^\star(A,X) = \left(\frac{I(A=1)}{\hat{e}(A=1|X)} - \frac{I(A=0)}{\hat{e}(A=0|X)}\right)(Y - \hat{\mu}(A,X)),$$

which is precisely the part of the influence function that depends on μ.

There are advantages in using exactly the same procedure in Example 3 for bounded continuous outcomes in order to force the target to satisfy also global constraints of P_0. We refer to Chapter 7 Van der Laan et al. (2011).

3.3 Algorithmic Procedure for TMLE

We now describe the algorithmic procedure of TMLE.

- Split data in two samples of size n_1 and n_2;
- With first sample, find the initial model:

$$\mu_{\hat{P}_{n_1}} = \arg\min_{\mu} \frac{1}{n_1} \sum_i (Y_i - \mu(A_i, X_i))^2;$$

- With the second sample, solve for the parametric submodel using some proposal function L (here we use L2):

$$\epsilon^\star = \arg\min \frac{1}{n_2} \sum_{i=1}^{n} \left(Y_i - \mu_{\hat{P}_{n_1}}(A_i, X_i) - \epsilon H^\star(A_i, X_i)\right)^2;$$

- Compute the target step $\mu^\star = \mu_{\hat{P}_{n_1}} + \epsilon^\star H^\star$
- With the second sample, compute the plug-in estimator

$$\frac{1}{n_2} \sum_i \psi(Q^\star_{n_1}) = \frac{1}{n_2} \sum_i \left(\mu^\star(1, X_i) - \mu^\star(0, X_i)\right);$$

- Compute asymptotic valid confidence intervals: For each i in n_2 calculate

$$IF(D_i) = (\hat{\mu}(1,X_i) - \hat{\mu}(0,X_i)) + \left(\frac{I(A_i=1)}{e_{\hat{P}_n}(A_i=1|X_i)} - \frac{I(A_i=0)}{e_{\hat{P}_n}(A_i=0|X_i)}\right)(Y_i - \hat{\mu}(A_i,X_i)) - \psi(D_i).$$

  Then,

$$\sigma_{tmle} = \sqrt{\frac{1}{n_2} \sum_i IF^2(D_i)}.$$

Why TMLE?

- Plug-in estimator
- Stable: estimator do not blow up
- Flexible approach to leverage ML methods to answer causal queries
- Minimal assumptions
- Asymptotically efficient

References

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

M. J. Van der Laan, S. Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 10. Springer, 2011.