Chapter 6
Why TMLE?

Sherri Rose, Mark J. van der Laan

In the previous five chapters, we covered the targeted learning road map. This included presentation of the tools necessary to estimate causal effect parameters of a data-generating distribution. We illustrated these methods with a simple data structure: $O = (W, A, Y) \sim P_0$. Our target parameter for this example was $\Psi(P_0) = E_{W,0}[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)]$, which represents the causal risk difference under causal assumptions.

Throughout these chapters, the case for TMLE using super learning is compelling, but many of its properties have not been fully discussed, especially in comparison to other estimators. This chapter makes a comprehensive case for TMLE based on statistical properties and compares TMLE to maximum-likelihood-based substitution estimators of the g-formula (MLE) and estimating-equation-based methodology. We continue to refer to the simple data structure $O = (W, A, Y) \sim P_0$ and causal risk difference as the target parameter in some comparisons, but also discuss the performance of TMLE and other estimators globally, considering many target parameters and data structures.

As we introduced in Chaps. 4 and 5, TMLE has many attractive properties that make it preferable to other existing procedures for estimation of a target parameter of a data-generating distribution for arbitrary semiparametric statistical models. TMLE removes all the asymptotic residual bias of the initial estimator for the target parameter if it uses a consistent estimator of the treatment mechanism. If the initial estimator is already consistent for the target parameter, the minimal additional fitting of the data in the targeting step may potentially remove some finite sample bias and certainly preserve this consistency property of the initial estimator. As a consequence, TMLE is a so-called double robust estimator.

In addition, if the initial estimator and the estimator of the treatment mechanism are both consistent, then it is also asymptotically efficient according to semiparametric statistical model efficiency theory. That is, under this condition, other competing estimators will underperform in comparison for large enough sample sizes with respect to variance, assuming that the competitors are required to have a bias for the target parameter smaller than $1/\sqrt{n}$ across a neighborhood of distributions of the

M.J. van der Laan and S. Rose, *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, DOI 10.1007/978-1-4419-9782-1_6,

true P_0 that shrinks to P_0 at this same rate $1/\sqrt{n}$. It allows the incorporation of machine learning (i.e., super learning) methods for the estimation of both the relevant part of P_0 and the nuisance parameter g_0 required for the targeting step, so that we do not make assumptions about the probability distribution P_0 we do not believe. In this manner, every effort is made to achieve minimal bias and the asymptotic semiparametric efficiency bound for the variance. We further explain these issues in the pages that follow.

Portions of this chapter are technical, but a general understanding of the essential concepts can be gleaned from reading the introduction to each of the sections and the tables at the end of each section. For example, Sect. 6.1 explains that there are two general types of estimators and provides a list of various estimators that may be familiar to the reader. Similarly, Sects. 6.2–6.6 discuss properties of TMLE: it is a loss-based, well-defined, unbiased, efficient substitution estimator of target parameters of a data-generating distribution. The introductions explain these concepts and the closing tables summarize these properties among competing estimators and TMLE. Therefore, a strong math background is not required to understand the basic concepts, and some readers may find it useful to skim or skip certain subsections.

6.1 Landscape

In order to effectively establish the benefits of TMLE, we must enumerate competing estimators. For example, what are our competitors for the estimation of causal effect parameters, such as $E_0Y_1 - E_0Y_0$, as well as other target parameters? We group these estimators into two broad classes: MLE and estimating equation methodology. For each specific estimation problem, one can come up with a number of variations of an estimator in such a class. In Chaps. 7 and 21, among others, we provide a finite sample comparison of TMLE with a number of estimators, including estimators specifically tailored for this simple data structure. Recall that the conditional expectation of Y given (A, W) is denoted $E_0(Y \mid A, W) \equiv \bar{Q}_0(A, W)$. Additionally, we let $Q_n = (\bar{Q}_n, Q_{W,n})$ be the estimate of the conditional mean and the empirical distribution for the marginal distribution of W, representing the estimator of the true $Q_0 = (\bar{Q}_0, Q_W)$.

6.1.1 MLE

A maximum likelihood estimator for a parametric statistical model $\{p_\theta : \theta\}$ is defined as a maximizer over all densities in the parametric statistical model of the empirical mean of the log density:

$$\theta_n = \arg\max_\theta \sum_{i=1}^{n} \log p_\theta(O_i).$$

The $L(p)(O) = -\log p(O)$ is called a loss function at candidate density p for the true density p_0 since its expectation is minimized across all densities p by the true density $p = p_0$. This minimization property of the log-likelihood loss function is the principle behind maximum likelihood estimation providing the basis for establishing that maximum likelihood estimators for correctly specified statistical models approximate the true distribution P_0 for large sample size.

An estimator that is based on maximizing the log-likelihood over the whole statistical model or submodels of the statistical model or utilizes algorithms that involve maximization of the log-likelihood will be called a maximum-likelihood-based estimator. We use the abbreviation MLE to refer specifically to maximum-likelihood-based substitution estimators of the g-formula.

This chapter can be equally applied to the case where $L(p)(O)$ is replaced by any other loss function $L(Q)$ for a relevant part Q_0 of p_0, satisfying that $E_0 L(Q_0)(O) \leq E_0 L(Q)(O)$ for each possible Q. In that case, we might call this estimator a minimum-loss-based estimator. TMLE incorporates this case as well, in which it could be called targeted minimum-loss-based estimation (still abbreviated as TMLE). In this chapter we focus our comparison on the log-likelihood loss function and will thereby refer to MLE, including ML-based super learning.

The g-formula was previously discussed in Chaps. 1–4. Recall that uppercase letters represent random variables and lowercase letters are a specific value for that variable. $\Psi(P_0)$ for the causal risk difference can be written as the g-formula:

$$\Psi(P_0) = \sum_w \Big[\sum_y y P_0(Y = y \mid A = 1, W = w) - \sum_y y P_0(Y = y \mid A = 0, W = w) \Big] P_0(W = w), \tag{6.1}$$

where

$$P_0(Y = y \mid A = a, W = w) = \frac{P_0(W = w, A = a, Y = y)}{\sum_y P_0(W = w, A = a, Y = y)}$$

is the conditional probability distribution of $Y = y$, given $A = a$, $W = w$, and

$$P_0(W = w) = \sum_{y,a} P_0(W = w, A = a, Y = y).$$

Recall that our target parameter only depends on P_0 through the conditional mean $\bar{Q}_0(A, W) = E_0(Y \mid A, W)$ and the marginal distribution Q_W of W; thus we can also write $\Psi(Q_0)$.

Maximum-likelihood-based substitution estimators of the g-formula are obtained by substitution of a maximum-likelihood-based estimator of Q_0 into the parameter mapping $\Psi(Q_0)$. The marginal distribution of W can be estimated with the nonparametric maximum likelihood estimator, which happens to be the empirical distribution that puts mass $1/n$ on each W_i, $i = 1, \ldots, n$. In other words, we estimate the expectation over W with the empirical mean over W_i, $i = 1, \ldots, n$. Maximum-

likelihood-based estimation of $\bar{Q}_0$ can range from the use of stratification to super learning. We introduced nonparametric estimation of $\bar{Q}_0$ in Chap. 3. Maximum-likelihood-based substitution estimators will be of the type

$$\psi_n = \Psi(Q_n) = \frac{1}{n}\sum_{i=1}^{n}\{\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)\}, \tag{6.2}$$

where this estimate is obtained by plugging in $Q_n = (\bar{Q}_n, Q_{W,n})$ into the mapping Ψ.

MLE using stratification. The simplest maximum likelihood estimator of $\bar{Q}_0$ stratifies by categories or possible values for (A, W). One then simply averages across the many categories (also called bins or treatment/covariate combinations). In most data sets, there will be a large number of categories with few or zero observations. One might refer to this as the curse of dimensionality, making the MLE for nonparametric statistical models typically ill defined, and an overfit to the data resulting in poor finite sample performance. One can refer to this estimator as the nonparametric MLE (NPMLE).

MLE after dimension reduction: propensity score methods. To deal with the curse of dimensionality, one might propose a dimension reduction W^r of W and apply the simple MLE to the reduced-data structure (W^r, A, Y). However, such a dimension reduction could easily result in a biased estimator of $\Psi(Q_0)$ by excluding confounders. One can show that a sufficient confounder is given by the propensity score $g_0(1 \mid W) = P_0(A = 1 \mid W)$, allowing one to reduce the dimension of W to only a single covariate, without inducing bias. A maximum likelihood estimator of $E_0(Y \mid A, W^r)$ can then be applied, where $W^r = g_0(1 \mid W)$, using stratification. For example, one creates five categories for the propensity score, thereby creating a total of ten categories for (A, W^r), and estimates $E_0(Y \mid A, W^r)$ with the empirical average of the outcomes within each category. Of course, this propensity score is typically unknown and will thus first need to be estimated from the data.

MLE using regression in a parametric working model. $\bar{Q}_0(A, W)$ is estimated using regression in a parametric working (statistical) model and plugged into the formula given in (6.2).

ML-based super learning. We estimate $\bar{Q}_0$ with the super learner, in which the collection of estimators may include stratified maximum likelihood estimators, maximum likelihood estimators based on dimension reductions implied by the propensity score, and maximum likelihood estimators based on parametric working models, beyond many other machine learning algorithms for estimation of $\bar{Q}_0$. Super learning requires a choice of loss function. If the loss function is a log-likelihood loss, $L(P_0)(O) = -\log p_0(O)$, then we would call this maximum-likelihood-based super learning. However, one might use a loss function for the relevant part $\bar{Q}_0$ that is not necessarily a log-likelihood loss, in which case we should call it minimum-loss-based super learning. For example, if Y is a continuous random variable with outcomes in $[0, 1]$, then one can select as loss function for $\bar{Q}_0$ the following function:

$$L(\bar{Q}_0)(O) = -Y \log \bar{Q}_0(A, W) + (1 - Y) \log(1 - \bar{Q}_0(A, W)),$$

which indeed satisfies that the expectation $E_0 L(\bar{Q})(O)$ is minimized by $\bar{Q} = \bar{Q}_0$. This loss function is an example of a loss function that is not a log-likelihood loss function. In this chapter we will not stress this additional important gain in generality of loss-based super learning relative to maximum-likelihood-based estimation, allowing us to proceed directly after the relevant parts of the distribution of P_0 required for evaluation of our target parameter $\Psi(P_0)$.

6.1.2 Estimating Equation Methods

Estimating-equation-based methodology for estimation of our target parameter $\Psi(P_0)$ includes inverse probability of treatment-weighted (IPTW) estimators and augmented IPTW (A-IPTW) estimators. These methods aim to solve an estimating equation in candidate ψ-values. An estimating function is a function of the data O and the parameter of interest. If $D(\psi)(O)$ is an estimating function, then we can define a corresponding estimating equation:

$$0 = \sum_{i=1}^{n} D(\psi)(O_i),$$

and solution ψ_n satisfying $\sum_{i=1}^{n} D(\psi_n)(O_i) = 0$. Most estimating functions for ψ will also depend on an unknown "nuisance" parameter of P_0. So we might define the estimating function as $D(\psi, \eta)$, where η is a candidate for the nuisance parameter. Given an estimator η_n of the required true nuisance parameter η_0 of P_0, we would define the estimating equation as

$$0 = \sum_{i=1}^{n} D(\psi, \eta_n)(O_i),$$

with solution ψ_n satisfying $\sum_{i=1}^{n} D(\psi_n, \eta_n)(O_i) = 0$. The theory of estimating functions teaches us that for each semiparametric statistical model and each target parameter, a class of estimating functions can be mathematically derived in terms of the gradients of the pathwise derivative of the target parameter, and the optimal estimating function that may yield an estimator with minimal asymptotic variance needs to be defined by the efficient influence curve (also called canonical gradient of the pathwise derivative) of the target parameter.

When the notation $D^*(\psi_0, \eta_0)$ is used for the estimating function $D(\psi_0, \eta_0)$, $D^*(\psi_0, \eta_0)$ is an estimating function implied by the efficient influence curve. An efficient influence curve is $D^*(P_0)(O)$, i.e., a function of O, but determined by P_0, and may be abbreviated $D^*(P_0)$ or $D^*(O)$. An optimal estimating function is one such that $D(\psi_0, \eta_0) = D^*(P_0)$.

For estimation of the causal risk difference, the following are two popular examples of estimating-equation-based methods, where the A-IPTW estimator is based on the estimating function implied by the efficient influence curve.

IPTW. One estimates our target parameter, the causal risk difference $\Psi(P_0)$, with

$$\psi_n = \frac{1}{n}\sum_{i=1}^{n}\{I(A_i = 1) - I(A_i = 0)\}\frac{Y_i}{g_n(A_i, W_i)}.$$

This estimator is a solution of an IPTW estimating equation that relies on an estimate of the treatment mechanism, playing the role of a nuisance parameter of the IPTW estimating function.

A-IPTW. One estimates $\Psi(P_0)$ with

$$\psi_n = \frac{1}{n}\sum_{i=1}^{n}\frac{\{I(A_i = 1) - I(A_i = 0)\}}{g_n(A_i, W_i)}(Y_i - \bar{Q}_n(A_i, W_i)) + \frac{1}{n}\sum_{i=1}^{n}\{\bar{Q}_n(1, W_i) - \bar{Q}_n(0, W_i)\}.$$

This estimator is a solution of the A-IPTW estimating equation that relies on an estimate of the treatment mechanism g_0 and the conditional mean $\bar{Q}_0$. Thus $(g_0, \bar{Q}_0)$ plays the role of the nuisance parameter of the A-IPTW estimating function. The A-IPTW estimating function evaluated at the true $(g_0, \bar{Q}_0)$ and true ψ_0 actually equals the efficient influence curve at the true data-generating distribution P_0, making it an optimal estimating function.

6.2 TMLE is Based on (Targeted) Loss-Based Learning

Suppose one is given a loss function $L()$ for a parameter $Q_0 = Q(P_0)$ while the estimand ψ_0 of interest is determined by Q_0. Thus, $Q_0 = \arg\min_Q E_0 L(Q)(O)$, where the minimum is taken over all possible parameter values of Q. One can proceed by defining a collection of candidate estimators $\hat{Q}_k$ that map the data P_n into an estimate of Q_0, where such estimators can be based on aiming to minimize the expected loss $Q \to E_0 L(Q)(O)$. This family of estimators can be used as a library of the loss-based super learner, which will use cross-validation to determine the best weighted combination of all these candidate estimators. The resulting super learner estimate Q_n can now be mapped into the estimate $\Psi(Q_n)$ of the estimand ψ_0.

Such estimators have the following properties. Firstly, these estimators are generally well defined by being based on minimizing empirical risk and cross-validated risk with respect to the loss function $L()$ over the statistical model. Secondly, by definition, these substitution estimators fully respect the global constraints implied by the statistical model and the target parameter mapping Ψ. Thirdly, such estimators can incorporate the state of the art in machine learning. Fourthly, the loss function

$L(Q)$ can be selected to result in good estimators of the estimand ψ_0: in particular, the TMLE chooses a loss function and a cleverly chosen parametric working model to construct a targeted loss function whose empirical risk represents the fit of the TMLE. Finally, such estimators can be constrained to also solve a particular estimating equation that might be considered to yield advantageous statistical properties of the substitution estimator of $\Psi(Q_n)$: The TMLE enforces such a constraint by iteratively minimizing the empirical risk over the parametric working model through the current initial estimate.

6.2.1 Competitors

MLE is a loss-based learning methodology based on the log-likelihood loss function $L(P_0) = -\log P_0$. This explains many of the popular properties of maximum-likelihood-based estimation. Since the log-likelihood loss function measures the performance of a candidate probability distribution as a whole, it does not represent a targeted loss function when the parameter of interest is a small feature of P_0. The lack of targeting of the MLE is particularly apparent when the data structure O is high dimensional and the statistical model is large.

An estimating equation method (e.g., A-IPTW) is not a loss-based learning method. It takes as input not a particular loss function but an estimating function, and the estimator is defined as a solution of the corresponding estimating equation. The estimating function is derived from local derivatives of the target parameter mapping and thereby ignores the global constraints implied by the statistical model and by the target parameter mapping. These global constraints are important to put a natural brake on estimators, so that it is no surprise that estimating equation methods are often notoriously unstable under sparsity.

6.2.2 TMLE

TMLE (targeted minimum-loss-based estimation) is a targeted-loss-based learning methodology. It is targeted by its choice of loss function $L()$ and by the targeted minimization over cleverly chosen parametric working models through an initial estimate. The TMLE is driven by the *global* choices of the loss function and parametric working model, and not defined by its consequence that it solves the efficient influence curve estimating equation, as implied by the *local* derivative condition. For example, consider the data structure $O = (W, A, Y)$, with Y continuous and bounded between 0 and 1.

Suppose that the statistical model is nonparametric and that the estimand is the additive treatment effect $E_{W,0}[E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W)]$, as in our mortality example. To define a TMLE we could select the squared error loss function $L(\bar{Q})(O) = (Y - \bar{Q}_0(A, W))^2$ for the conditional mean $\bar{Q}_0$, and the linear parametric

working model $\bar{Q}(\epsilon) = \bar{Q} + \epsilon H^*$. Alternatively, we could define a TMLE implied by the "quasi"-log-likelihood loss function $-Y \log \bar{Q}_0(A, W) - (1-Y)\log(1-\bar{Q}_0(A, W))$, and the logistic linear parametric working model $\text{logit}\bar{Q}(\epsilon) = \text{logit}\bar{Q} + \epsilon H^*$.

Both TMLEs solve the efficient influence curve estimating equation, but they have very different properties regarding utilization of the global constraints of the statistical model. The TMLE with the squared error loss does not respect that it is known that $P_0(0 < Y < 1) = 1$, and, as a consequence, the TMLE $\bar{Q}_n^*$ can easily predict far outside [0, 1], making it an unstable estimator under sparsity. In fact, this TMLE violates the very principle of TMLE in that TMLE should use a parametric *submodel* through the initial estimator, and the linear fluctuations of an initial estimator $\bar{Q}_n^0$ do *not* respect that $0 < \bar{Q}_0 < 1$, and are thus not a *submodel* of the statistical model. On the other hand, the other *valid* TMLE uses a logistic fluctuation of the initial estimator that fully respects this constraint, and is therefore a sensible substitution estimator fully respecting the global constraints of the statistical model. We refer to Chap. 7 for a full presentation of the latter TMLE for continuous and bounded Y.

Table 6.1 Summary of loss-based estimators for a general $\Psi(P_0)$

	MLE					**Estimating equations**	
	TMLE	Stratification	Propensity score	Parametric regression	ML-based super learning	IPTW	A-IPTW
Loss-based estimator of $\Psi(P_0)$	×	×	×	×	×		

6.3 TMLE Is Well Defined

An estimator that is well defined is desirable. Well-defined estimators have one solution in the space of possible solutions. It is easy to see why a well-defined estimator would be preferable to one that is not well defined. We seek the best estimate of $\Psi(P_0)$, and if our estimator gives multiple or no solutions, that presents a problem.

6.3.1 Competitors

MLEs aim to maximize a log-likelihood over candidate parameter values. Thus, MLE is often well defined, since, even if there are local maxima, the empirical log-likelihood or cross-validated log-likelihood can be used to select among such local maxima. Estimating equation methods are not well defined in general since the only criterion is that it solves the equation. A maximum likelihood estimator in a para-

metric statistical model often cannot be uniquely defined as a solution of the score equation since each local maximum will solve the score equation. The estimating equation methods are well defined for our target parameter with the simple data structure $O = (W, A, Y)$ and nonparametric statistical model for P_0, as is obvious from the definition of the IPTW and A-IPTW estimators given above. This is due to the fact that the estimating functions happen to be linear in ψ, allowing for a simple closed-form solution to their corresponding estimating equations.

When defining an estimator as a solution of the optimal efficient score/influence curve estimating equation, one may easily end up having to solve nonlinear equations that can have multiple solutions. The estimating equation itself provides no information about how to select among these candidate estimates of the target parameter. Also, one cannot use the likelihood since these estimators cannot be represented as $\hat{\Psi}(P_n)$ for some candidate P_n, i.e., these solutions ψ_n of the estimating equation are not substitution estimators (Sect. 6.6). This goes back to the basic fact that estimating functions (such as those defined by the efficient score/efficient influence curve) might not asymptotically identify the target parameter, and, even if they did, the corresponding estimating equation might not uniquely identify an estimator for a given finite sample.

In addition, for many estimation problems, the efficient influence curve $D^*(P_0)$ of the target parameter cannot be represented as an estimating function $D^*(\psi_0, \eta_0)$, so that the estimating equation methodology is not directly applicable. This means that the estimating equation methodology can only be applied if the efficient influence curve allows a representation as an estimating function. This is not a natural requirement, since the efficient influence curve $D^*(P_0)$ is defined as a gradient of the pathwise derivative of the target parameter along paths through P_0, and thereby only defines it as a function of P_0. There is no natural reason why the dependence of $D^*(P_0)$ on P_0 can be expressed in a dependence on two variation-independent parameters (ψ_0, η_0). Indeed, in some of our chapters we encounter target parameters where the efficient influence curve does not allow a representation as an estimating function.

6.3.2 TMLE

Unlike estimating function methodology (e.g., A-IPTW), TMLE does not aim to solve an estimating equation but instead uses the log-likelihood as a criterion. The super learner, representing the initial estimator in the TMLE, uses the (cross-validated) log-likelihood, or other loss function, to select among many candidate estimators. Even in the unlikely event that more than one global maximum exists, both would provide valid estimators so that a simple choice could make the super learner well defined. The targeting step involves computing a maximum likelihood estimator in a parametric working model of the same dimension as the target parameter, fluctuating the initial estimator, and is therefore as well defined as a parametric maximum likelihood estimator; again, the log-likelihood can be used to select

among different local maxima. See Table 6.2 for a summary of well-defined estimators of $\Psi(P_0)$.

Table 6.2 Summary of well-defined estimators for a general $\Psi(P_0)$

	MLE					**Estimating equations**	
	TMLE	Stratification	Propensity score	Parametric regression	ML-based super learning	IPTW	A-IPTW
Well-defined estimator of $\Psi(P_0)$	×	×	×	×	×		

6.4 TMLE Is Unbiased

An estimator is asymptotically unbiased if it is unbiased as the sample size approaches infinity. Bias is defined as follows: $\text{bias}(\psi_n) = E_0(\psi_n) - \psi_0$, where $E_0(\psi_n)$ denotes the expectation of the estimator ψ_n viewed as a function of the n i.i.d. copies $O_1, \ldots, O_n$ drawn from P_0. An estimator is unbiased if $\text{bias}(\psi_n) = 0$. It is rare that an estimator is exactly unbiased. If we restricted ourselves to using only unbiased estimators, then in most estimation problems we would have no estimators available. Therefore, one wants to focus on estimators where bias is negligible for the purpose of obtaining confidence intervals for ψ_0 and tests of null hypotheses about ψ_0. This can be achieved by requiring that the bias converge to zero when sample size n increases, at a rate smaller than $1/\sqrt{n}$, such as $1/n$. Indeed, most correctly specified parametric maximum likelihood estimators have a bias of the order $1/n$.

Why do we care about bias? In the real world, biased estimators can lead to false positives in multimillion-dollar studies. That is, the true causal risk difference might be equal to zero, but if the estimator is biased, then a test that ignores this bias will interpret the bias as a deviation from the null hypothesis. This deviation from the null hypothesis would be declared statistically significant if sample size was large enough. In addition, bias against the null hypothesis (for example, one wishes to test for a positive treatment effect, but the effect estimate is biased low) results in less power to reject the null hypothesis. Overall, bias causes incorrect statistical inference.

One might wonder why one would not aim to estimate the bias of an estimator. The problem is that estimation of bias is typically an impossible goal, inducing more error than the bias: often the best one can do is to diagnose the presence of unusual bias, and that is indeed a task that should be incorporated in a data analysis (Chap. 10). Again, our goal is the best estimator of the true effect, and an asymptotically biased estimator is an estimator that cannot even learn the truth. We also want the bias to be asymptotically negligible so that statistical assessment of uncertainty based on an estimator of the variance of the estimator is reasonably valid.

6.4.1 Competitors

MLEs using stratification, super learning, or parametric regression are asymptotically unbiased if $\bar{Q}_0$ is consistently estimated. In order for propensity score methods that fit a nonparametric regression on treatment A and the propensity score to be asymptotically unbiased, the estimator of g_0 must be consistent. MLEs using stratification can easily suffer from large finite sample bias in sparse data. In other words, using a nonparametric MLE with a limited data set provides no recipe for an unbiased estimator of the target parameter $\Psi(P_0)$. IPTW is asymptotically unbiased for $\Psi(P_0)$ if the estimator of g_0 is consistent, and A-IPTW is asymptotically unbiased for $\Psi(P_0)$ if either $\bar{Q}_0$ or g_0 is consistently estimated. The asymptotic bias of the A-IPTW is characterized by the same expression provided in the next paragraph for TMLE. The finite sample bias is very much a function of how g_0 is estimated, in particular with respect to what covariates are included in the treatment mechanism and how well it approximates the true distribution. Since g_n is by necessity estimated based on the log-likelihood for the treatment mechanism, its fit is not affected by data on Y. As a consequence, covariates that have no effect on Y but a strong effect on A will be included, only harming the bias reduction effort.

6.4.2 TMLE

Using super learning within TMLE makes our estimator of the outcome regression $\bar{Q}_0$ and estimator of the treatment mechanism g_0 maximally asymptotically unbiased. In our flexible nonparametric statistical model, we can show that the asymptotic bias in our procedure involves a product of the bias of $\bar{Q}_n^*$ and g_n relative to the true $\bar{Q}_0$ and g_0, respectively. For example, with data structure $O = (W, A, Y)$ in an observational study (where g_0 is unknown), our asymptotic bias of the TMLE $\Psi(Q_n^*)$ given by

$$\text{bias}(\psi_n) = P_0\Big\{\frac{g_0(1 \mid W) - g(1 \mid W)}{g(1 \mid W)}(\bar{Q}_0 - \bar{Q}^*)(1, W) - \frac{g_0(0 \mid W) - g(0 \mid W)}{g(0 \mid W)}(\bar{Q}_0 - \bar{Q}^*)(0, W)\Big\},$$

where $\bar{Q}^*$ and g denote the limits of $\bar{Q}_n^*$ and g_n. This teaches us that the asymptotic bias behaves as a second-order difference involving the product of approximation errors for g_0 and $\bar{Q}_0$. The empirical counterpart of this term plays the role of second-order term for the TMLE approximation of the true ψ_0, and thereby also drives the finite sample bias. For reliable confidence intervals one wants $\sqrt{n}$ times the empirical counterpart of this bias term to converge to zero in probability as sample size converges to infinity. If one wants to make this second-order term and the resulting bias as small as possible, then theory teaches us that we should use super learning

for both $\bar{Q}_0$ and g_0. As we point out in the next subsection, to minimize the variance of the first-order mean zero linear approximation of the TMLE approximation of the true ψ_0, one needs to estimate $\bar{Q}_0$ consistently. In other words, the use of super learning is essential for both maximizing efficiency as well as minimizing bias. For a formal theorem formalizing these statements we refer the interested reader to Chap. 27.

From this bias term one concludes that if the estimator of g_0 is correct, our estimator will have no asymptotic bias. This is an important scenario: consider again $O = (W, A, Y)$, and suppose we know the treatment mechanism, such as an RCT. In this instance, TMLE is always unbiased. Additionally, finite sample bias can be removed in RCTs by estimating g_0. If $\bar{Q}_n^*$ is already close to $\bar{Q}_0$, then the targeting step will further reduce the bias if g_n is also consistent. Finally, since running an additional univariate regression of the clever covariate on the outcome using the initial estimator as offset is a robust operation (assuming the clever covariate is bounded), even if g_n is misspecified, the targeting step will not cause harm to the bias.

In fact, one can show that if one replaces $g_0(A \mid W)$ by a true (sufficient) conditional distribution g_0^s of A, given a subset W^s of all covariates W, and W^s is chosen such that $Q^* - Q_0$ only depends on W through W^s, then the TMLE using this g_0 is also an unbiased estimator of the estimand ψ_0. Here Q^* represents the possibly misspecified estimand of the TMLE $\bar{Q}_n^*$. That is, the TMLE already achieves its full bias reduction by only incorporating the covariates in the treatment mechanism that explain the residual bias of $\bar{Q}_n^*$ with respect to $\bar{Q}_0$. We say that the TMLE is collaborative double robust to stress that consistency of the TMLE of ψ_0 is already achieved if g_n appropriately adapts to the residual bias of $\bar{Q}_n^*$: the TMLE is collaborative double robust, which is a stronger type of robustness with respect to misspecification of the nuisance parameters $\bar{Q}_0$ and g_0 than double robustness. In particular, an estimator g_n of g_0 used by the TMLE does not need to include covariates that are not predictive of Y, and are thus not confounders, even if the true treatment mechanism used these covariates. Apparently, the selection of covariates to be included in the estimator of the treatment mechanism should not be based on how well it fits g_0, but on the gain in fit of $\bar{Q}_0$ obtained by fitting the parametric working model (that uses this estimate of g_0) through the initial estimator $\bar{Q}_n^0$, relative to the fit of the initial estimator.

That is, TMLE naturally allows for the fine-tuning of the choice of g_n based on the fit of the corresponding TMLE of $\bar{Q}_0$, and can thereby data-adaptively select covariates into the treatment mechanism that actually matter and yield effective bias reduction in the TMLE step. For example, consider two possible estimators g_n^1 and g_n^2. These two choices combined with the initial estimator $\bar{Q}_n^0$ yield two different TMLEs, $\bar{Q}_{n1}^*$ and $\bar{Q}_{n2}^*$. These results suggest that one should select the estimator of g_0 for which the TMLE has the best fit of $\bar{Q}_0$. Note that this is equivalent to selecting covariates for the treatment mechanism based on how well the resulting estimate of the treatment mechanism improves the predictiveness of the corresponding clever covariate in predicting the outcome Y beyond the initial regression. This insight that the choice of g_n should be based on an evaluation of the resulting TMLE of $\bar{Q}_0$ is formalized by collaborative TMLE (C-TMLE), which is presented in Chaps. 19–21

and 23. See Table 6.3 for a summary of conditions for unbiased estimation among the estimators for a general $\Psi(P_0)$. Table 6.4 summarizes targeted estimation of the treatment mechanism for a general $\Psi(P_0)$.

Table 6.3 Summary of conditions for unbiased estimation for a general $\Psi(P_0)$

	MLE					**Estimating equations**	
	TMLE	Stratification	Propensity score	Parametric regression	ML-based super learning	IPTW	A-IPTW
Consistent estimation of $\bar{Q}_0$		×		×	×		
Consistent estimation of g_0			×			×	
Consistent estimation of $\bar{Q}_0$ or g_0	×						×
Problems in finite samples		×					

Table 6.4 Summary of targeted estimation of the treatment mechanism for a general $\Psi(P_0)$

	MLE					**Estimating equations**	
	(C-)TMLE	Stratification	Propensity score	Parametric regression	ML-based super learning	IPTW	A-IPTW
Targeted estimation of treatment mechanism	×						

6.5 TMLE Is Efficient

Efficiency is another measure of the desirability of an estimator. Finite sample efficiency for an estimator ψ_n can be defined as

$$\text{efficiency}(\psi_n) = \frac{\left(\frac{1}{I(\Psi(P_0))}\right)}{n\text{var}(\psi_n)},$$

where $I(\Psi(P_0))$ is the Fisher information, defined as 1 over the variance of the efficient influence curve. The variance of the efficient influence curve is also called the generalized Cramer–Rao lower bound for the variance of locally (approximately) unbiased estimators. Thus, efficiency(ψ_n) is the ratio of the minimum possible asymptotic variance for an approximately unbiased estimator over its actual finite sample variance. The asymptotic efficiency is defined as the limit of efficiency(ψ_n) for n converging to infinity. If the estimator of $\Psi(P_0)$ is unbiased and the asymptotic efficiency(ψ_n) = 1, the estimator is asymptotically efficient. Asymptotically efficient estimators achieve the Cramer–Rao bound (i.e., the variance of an unbiased estimator is, at a minimum, the inverse of the Fisher information) for large n. What we really care about, though, is performance in finite samples. So we would like to see that the finite sample efficiency efficiency(ψ_n) is close to 1. Minimally, we want an asymptotically efficient estimator, but we also want our estimator to perform well in realistic finite sample sizes.

Efficiency theory is concerned with an admission criterion: it is restricted to only those estimators that have negligible bias (i.e., small bias in finite samples) along small fluctuations of the true data-generating distribution, and among such estimators it defines a best estimator as the estimator that has the smallest asymptotic variance. This best estimator will be asymptotically linear with influence curve the efficient influence curve $D^*(O)$. An estimator $\hat{\Psi}(P_n)$ of ψ_0 is asymptotically linear with influence curve $IC(O)$ if it satisfies

$$\sqrt{n}(\psi_n - \psi_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IC(O_i) + o_{P_0}(1).$$

Here the remainder term, denoted by $o_{P_0}(1)$, is a random variable that converges to zero in probability when the sample size converges to infinity. Asymptotic linearity is a desirable property as it indicates that the estimator behaves like an empirical mean, and, as a consequence, its bias converges to zero in sample size at a rate faster than $1/\sqrt{n}$, and, for n large enough, it is approximately normally distributed. The influence curve of an estimator evaluated as a function in O measures how robust the estimator is toward extreme values. The influence curve $IC(O)$ has a mean of zero under sampling from the true probability distribution P_0, and its (finite) variance is the asymptotic variance of the standardized estimator $\sqrt{n}(\psi_n - \psi_0)$. In other words, the variance of $\hat{\Psi}(P_n)$ is well approximated by the variance of the influence curve, divided by sample size n. An estimator is asymptotically efficient if and only if its influence curve is equal to the efficient influence curve, $IC(O) = D^*(O)$.

If we already agree that we want unbiased estimators, why do we care about efficiency? Given two unbiased estimators why should we choose the one that is also efficient? An unbiased estimator that has a large spread (i.e., huge confidence intervals) may be uninformative. A practical real-world result of this, aside from improved interpretation, is huge potential cost savings. If we can extract more information out of the our data with an efficient estimator, we can reduce the sample size required for an inefficient estimator. This savings may be nontrivial. For example, in

a large multicenter RCT with a projected budget of $100 million, reducing sample size by 30% results in close to $30 million saved.

6.5.1 Competitors

If the covariate W is discrete, MLE using stratification is efficient asymptotically, but falls apart in finite samples if the number of categories is large. Suppose we have 30 discrete covariates, each with 3 levels. This gives us 3^{30} different covariate combinations, over 200 trillion! It is clear it becomes hopeless to wish for efficiency in finite sample sizes.

If W also includes continuous components, and some form of smoothing is used in the maximum likelihood estimation of $\bar{Q}_0$, then the maximum likelihood estimator will have approximation errors of the form $\sum_w E(\bar{Q}_n(1,w) - \bar{Q}_0(1,w))P_0(W = w)$ (minus the same term with $A = 0$). That is, the bias of $\bar{Q}_n$ will translate directly into a bias for the substitution estimator, and this bias will typically not be $o_P(1/\sqrt{n})$. The bias will also be larger than it would have been using super learning. In these cases, the bias causes the MLE to not be asymptotically linear and thereby also not achieve asymptotic efficiency. As discussed above, TMLE reduces the bias into a second-order term, so that it can still be asymptotically linear and thus efficient when the MLE will not (e.g., if g_0 can be well estimated).

Estimating equation methodology using the optimal estimating function (implied by the efficient influence curve) is asymptotically efficient if both $\bar{Q}_n$ and g_n are estimated consistently and if these estimators approximate the truth fast enough so that the estimator of ψ_0 succeeds in being asymptotically linear. This would require using super learning to estimate the nuisance parameters of the optimal estimating function. Due to the fact that estimating-equation-based estimators are not substitution estimators (Sect. 6.6), these estimators ignore global constraints, which harms the finite sample efficiency, in particular in the context of sparsity.

6.5.2 TMLE

Like the optimal estimating equation based estimator (i.e., A-IPTW), TMLE is double robust and (locally) efficient under regularity conditions. In other words, if the second-order term discussed above is asymptotically negligible, then the TMLE is consistent and asymptotically linear if either $\bar{Q}_n$ or g_n is a consistent estimator, and if both estimators are asymptotically consistent, then the TMLE is asymptotically efficient. TMLE also has excellent finite sample performance because it is driven by a log-likelihood (or other loss function) criterion, and a substitution estimator respecting all global constraints. The finite sample efficiency is further enhanced by the natural potential to fine-tune the estimator of the treatment mechanism through the predictiveness of the corresponding clever covariate, so that the treatment mech-

anism can be fitted in a way that is beneficial to its purpose in the targeting step. As previously noted, this is formalized by C-TMLE considered in later chapters. See Table 6.5 for a summary of efficiency among estimators for a general $\Psi(P_0)$.

Table 6.5 Summary of efficiency among estimators for a general $\Psi(P_0)$

	MLE					**Estimating equations**	
	TMLE	Stratification	Propensity score	Parametric regression	ML-based super learning	IPTW	A-IPTW
Efficient estimator of $\Psi(P_0)$	×	×					×
Problems in finite samples		×	×			×	×

6.6 TMLE Is a Substitution Estimator

Substitution estimators can be written as a mapping, taking an estimator of the relevant part of the data-generating distribution (e.g., P_n, P_n^*, Q_n, Q_n^*) and plugging it into the mapping $\Psi()$. The substitution estimator respects the statistical model space (i.e., the global constraints of the statistical model). Knowing and using information about the global constraints of the statistical model is helpful for precision (efficiency), particularly in the context of sparsity. For example, a substitution estimator of the risk difference ψ_0 respects knowledge that the mean outcome regression $\bar{Q}_0$ is bounded between $[0, 1]$, or that ψ_0 is a difference of two probabilities.

To understand why respecting global constraints in a statistical model is important in the context of sparsity (i.e., the data carry little information for target parameter), suppose one wishes to estimate the mean of an outcome Y based on observing n i.i.d. copies $Y_1, \ldots, Y_n$. Suppose it is also known that $E_0 Y$ is larger than 0 and smaller than 0.1. This knowledge is not needed if the sample size is large enough such that the standard error of the estimator is much smaller than 0.1, but for small sample sizes, it cannot be ignored.

6.6.1 Competitors

MLEs using stratification, super learning, propensity scores, and parametric regression are substitution estimators. An estimator of ψ_0 that is obtained as a solution of an estimating equation is often *not* a substitution estimator, i.e., it cannot be written as $\Psi(P_n)$ for a specified estimator P_n of P_0 in the statistical model. Indeed, IPTW

and A-IPTW are not substitution estimators. To be specific, suppose one wishes to estimate the treatment-specific mean $E_0Y_1 = E_0[E_0(Y \mid A = 1, W)]$ based on n i.i.d. copies of (W, A, Y), Y being binary. In this case, the A-IPTW estimator ψ_n, which solves the efficient influence curve estimating equation, can fall outside the range $[0, 1]$, due to inverse probability of treatments being close to zero. This proves that it is not a substitution estimator, which results in a loss of finite sample efficiency.

6.6.2 TMLE

The TMLE of ψ_0 is obtained by substitution of an estimator P_n^* into the mapping $\Psi()$. For the risk difference, this mapping is given in (6.1). As a consequence, it respects the knowledge of the statistical model. TMLE for the treatment-specific mean, discussed above, would result in E_0Y_1 between $[0, 1]$. See Table 6.6 for a summary of substitution estimators for a general $\Psi(P_0)$.

Table 6.6 Substitution estimators for a general $\Psi(P_0)$

	MLE					**Estimating equations**	
	TMLE	Stratification	Propensity score	Parametric regression	ML-based super learning	IPTW	A-IPTW
Substitution estimator of $\Psi(P_0)$	×	×	×	×	×		

6.7 Summary

The TMLE procedure produces a well-defined, unbiased, efficient substitution estimator of target parameters of a data-generating distribution. Competing estimators, falling into the broad classes of MLE and estimating equation methodology, do not have all of these properties and will underperform in many scenarios in comparison to TMLE. See Table 6.7 for a summary of statistical properties among estimators for a general $\Psi(P_0)$.

6.8 Notes and Further Reading

We refer readers to the references listed in Chap. 4. Appendix A covers further theoretical development of TMLE. A key reference for propensity score methods is Rosenbaum and Rubin (1983), and we also refer readers to Chap. 21.

Table 6.7 Summary of statistical properties among estimators for a general $\Psi(P_0)$

	MLE					**Estimating equations**	
	TMLE	Stratification	Propensity score	Parametric regression	ML-based super learning	IPTW	A-IPTW
Loss-based:							
Loss-based estimator of $\Psi(P_0)$	×	×	×	×	×		
Well-defined:							
Well-defined estimator of $\Psi(P_0)$	×	×	×	×	×		
Unbiased under:							
Consistent estimation of $\bar{Q}_0$		×		×	×		
Consistent estimation of g_0			×			×	
Consistent estimation of $\bar{Q}_0$ or g_0	×						×
Problems in finite samples		×					
Efficiency:							
Efficient estimator of $\Psi(P_0)$	×	×					×
Problems in finite samples		×	×			×	×
Substitution estimator:							
Substitution estimator of $\Psi(P_0)$	×	×	×	×	×		