

One-Step Targeted Minimum Loss-based Estimation Based on Universal Least Favorable One-Dimensional Submodels

Mark J. van der Laan^{*}

Susan Gruber[†]

^{*}University of California, Berkeley, Division of Biostatistics, laan@berkeley.edu

[†]Harvard Medical School and Harvard Pilgrim Healthcare Institute, susan_gruber@harvardpilgrim.org

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

<http://biostats.bepress.com/ucbbiostat/paper347>

Copyright ©2016 by the authors.

One-Step Targeted Minimum Loss-based Estimation Based on Universal Least Favorable One-Dimensional Submodels

Mark J. van der Laan and Susan Gruber

Abstract

Consider a study in which one observes n independent and identically distributed random variables whose probability distribution is known to be an element of a particular statistical model, and one is concerned with estimation of a particular real valued pathwise differentiable target parameter of this data probability distribution. The targeted maximum likelihood estimator (TMLE) is an asymptotically efficient substitution estimator obtained by constructing a so called least favorable parametric submodel through an initial estimator with score, at zero fluctuation of the initial estimator, that spans the efficient influence curve, and iteratively maximizing the corresponding parametric likelihood till no more updates occur, at which point the updated initial estimator solves the so called efficient influence curve equation. In this article we construct a one-dimensional universal least favorable submodel for which the TMLE only takes one step, and thereby requires minimal extra data fitting to achieve its goal of solving the efficient influence curve equation. We generalize these to universal least favorable submodels through the relevant part of the data distribution as required for targeted minimum loss-based estimation. Finally, remarkably, given a multidimensional target parameter, we develop a universal canonical one-dimensional submodel such that the one-step TMLE, only maximizing the log-likelihood over a univariate parameter, solves the multivariate efficient influence curve equation. This allows us to construct a one-step TMLE based on a one-dimensional parametric submodel through the initial estimator, that solves any multivariate desired set of estimating equations.

1 Introduction

Targeted learning (van der Laan and Rubin, 2006; van der Laan, 2008; van der Laan and Rose, 2011) is a subfield of statistics concerned with the development of asymptotically efficient substitution estimators of specific target parameters of the data distribution, across possible data distributions within a realistic statistical model. By necessity any such procedure will have to integrate the state of the art in data adaptive estimation, but will also have to target such data adaptive estimators of relevant parts of the data distribution so that they are minimally biased for the target parameter.

Efficiency (Bickel et al., 1993) and empirical process theory (van der Vaart and Wellner, 1996) for general statistical models provide the foundation for the construction of such targeted machine learning algorithms. The canonical gradient of the pathwise derivative of the target parameter mapping defines an asymptotically efficient estimator with respect to the assumed statistical model as an estimator that is asymptotically linear with influence curve equal to the canonical gradient, which is the reason that the canonical gradient is also called the efficient influence curve. The construction of an efficient estimator of a pathwise differentiable target parameter will thereby naturally involve the utilization of this canonical gradient. The one-step estimator (e.g., (Bickel et al., 1993)) is such a general method that adds to an initial estimator of the target parameter the empirical mean of the estimated efficient influence curve. Estimating equation methodology (van der Laan and Robins, 2003; Robins and Rotnitzky, 1992) represents a related methodology that assumes that the efficient influence curve can be represented as an estimating function in the target parameter and a nuisance parameter, and defines the estimator as the solution of the resulting estimating equation. These procedures do not result in substitution estimators and thereby can lack finite sample robustness.

The targeted maximum likelihood estimator (TMLE) is a two-stage estimator obtained by constructing a parametric submodel through an initial estimator of the data distribution with score, at zero fluctuation of the initial estimator, that spans the efficient influence curve, and iteratively maximizing the corresponding parametric likelihood till no more updates occur. At that point the updated initial estimator solves the so called efficient influence curve equation (van der Laan and Rubin, 2006). The TMLE of the target parameter is now the corresponding plug-in estimator. The fact that the targeted estimator of the data distribution solves the efficient influence curve equation provides the basis for establishing the asymptotic efficiency of the TMLE under regularity conditions, beyond the crucial condition that the initial estimator is within a neighborhood (e.g., $n^{-1/4}$) of the true data distribution. To minimize the degree of violation of this crucial rate-of-convergence condition on the initial estimator as much as possible, we have proposed to construct such an initial estimator with the ensemble super-learner template fully utilizing the power and generality of cross-validation (van der Laan and Dudoit, 2003; van der Vaart et al.,

2006; van der Laan et al., 2006, 2007; Polley et al., 2012), while integrating the state of the art in machine learning. This super-learner has been proven to be optimal in the sense that it performs asymptotically as well as the best weighted combination of candidate estimators in its library of candidate estimators.

The parametric submodel through the initial estimator with a score that spans the efficient influence curve that is used in the TMLE procedure is called least favorable because it is the parametric submodel that maximizes the asymptotic variance of the submodel-specific maximum likelihood estimator of the target parameter under sampling from the initial estimator. In this article, we point out that this least favorable parametric submodel can also be interpreted as the submodel that maximizes the absolute infinitesimal change in target parameter (relative to initial estimator) divided by the information-norm of the infinitesimal change in probability distribution (relative to initial estimator). This provides a nice intuition about the targeted maximum likelihood step in TMLE as a fitting procedure that locally maximizes the change in target parameter per unit amount of fitting as measured by unit of information. However, it also shows that this choice of submodel is tailored to be optimal locally around the initial estimator, so that its ability to indeed provide maximal change in the target parameter per unit of information relies on the initial estimator being close enough to the true probability distribution.

This motivates us in this article to define and construct a one-dimensional universal least favorable submodel whose score equals the efficient influence curve at *each* of its parameter values, not just at 0. We show that such a universal least favorable submodel makes the targeted maximum likelihood estimator perform the desired job in one step, with minimal additional fitting of the data. As a consequence, it maximally preserves the statistical performance of the initial estimator, while achieving its desired targeted bias reduction. In particular, this universal least favorable submodel avoids the need for iterative targeted maximum likelihood estimation, and thereby possible overfitting in finite samples. It also provides the basis to various generalizations as needed for targeted minimum loss-based estimation of a possibly multivariate target parameter. Examples in the current literature in which the TMLE converged in one step happened to already use a universal least favorable submodel.

2 Statistical formulation of the goal and result of this article

Let O_1, \dots, O_n be n independent and identically distributed copies of a random variable $O \sim P_0$ with probability distribution P_0 that is known to be an element of a set \mathcal{M} of possible probability distributions. We refer to \mathcal{M} as the statistical model for the true data distribution P_0 . Let $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$ be a d -dimensional target parameter mapping, so that $\psi_0 = \Psi(P_0)$ represents the target parameter or estimand of interest that best approximates the answer to the question of interest.

We assume that Ψ is pathwise differentiable at each $P \in \mathcal{M}$ with canonical gradient $D^*(P)$. That is, for each path $\{P_{\epsilon,h} : \epsilon\}$ through P at $\epsilon = 0$ and score S_h , indexed by h in some index set \mathcal{H} , we have

$$\left. \frac{d}{d\epsilon} \Psi(P_{\epsilon,h}) \right|_{\epsilon=0} = PD^*(P)S_h,$$

where $Pf = \int f(o)dP(o)$ denotes the expectation operator w.r.t. P . $D^*(P)$ is the unique gradient that is also an element of the so called tangent space $T(P)$, defined as the closure of the linear span of all scores $\{S_h : h \in \mathcal{H}\}$ in the Hilbert space $L_0^2(P)$ of functions of O with mean zero under P , endowed with the inner-product $\langle S_1, S_2 \rangle = PS_1S_2$.

An estimator of ψ_0 is a mapping $\hat{\Psi}$ that maps the empirical probability distribution P_n of O_1, \dots, O_n into the parameter space $\Psi(\mathcal{M}) \subset \mathbb{R}^d$, and the corresponding estimate of ψ_0 is given by $\psi_n = \hat{\Psi}(P_n)$. An estimator $\hat{\Psi}(P_n)$ is asymptotically efficient at P_0 if and only if it is asymptotically linear with influence curve equal to the canonical gradient $D^*(P_0)$:

$$\hat{\Psi}(P_n) - \Psi(P_0) = (P_n - P_0)D^*(P_0) + o_P(1/\sqrt{n}).$$

Such an estimator satisfies (by CLT) $\sqrt{n}(\psi_n - \psi_0) \Rightarrow_d N(0, \Sigma_0 = P_0\{D^*(P_0)D^*(P_0)^\top\})$, so that statistical inference can be based on the estimator of its influence curve $D^*(P_0)$. The canonical gradient $D^*(P_0)$ of $\Psi : \mathcal{M} \rightarrow \mathbb{R}^d$ is also called the efficient influence curve.

A targeted maximum likelihood estimator (TMLE) is defined as follows. One first constructs an initial estimator $P_n^0 \in \mathcal{M}$ of P_0 . In addition, one defines a local least favorable parametric submodel $\{P_{n,\delta}^{0,\text{lflm}} : \delta\}$ through P_n^0 at $\delta = 0$ with d -dimensional parameter δ and with score $\left. \frac{d}{d\delta} \log dP_{n,\delta}^{0,\text{lflm}} / dP_n^0 \right|_{\delta=0} = D^*(P_n^0)$.

This is used to define the corresponding maximum likelihood estimator $\delta^0 = \arg \max P_n \log dP_{n,\delta}^{0,\text{lflm}} / dP_n^0$. The one-step TMLE of P_0 is now defined as $P_n^1 = P_{n,\delta^0}^{0,\text{lflm}}$. This process is iterated by defining $P_n^{k+1} = P_{n,\delta^k}^{k,\text{lflm}}$, $k = 1, 2, \dots$, till a $k = K$ for which $\delta^K \approx 0$. The TMLE of P_0 is then defined by the final update $P_n^* = P_{n,\delta^K}^{K,\text{lflm}}$, which solves $P_n D^*(P_n^*) \approx 0$. The TMLE of ψ_0 is the corresponding plug-in estimator $\Psi(P_n^*)$. Here ≈ 0 can be replaced by $o_P(1/\sqrt{n})$: for example, one might iterate till $\|P_n D^*(P_n^K)\| \leq 1/n$, where one could use the Euclidean norm. Below, we will ignore the numerical approximation error and just write $P_n D^*(P_n^*) = 0$.

The asymptotic efficiency of the TMLE, under regularity conditions, is established as follows. First, define the second order term $R_2(P, P_0)$ by the equation $\Psi(P) - \Psi(P_0) = (P - P_0)D^*(P) + R_2(P, P_0)$. Due to $D^*(P)$ being a canonical gradient, $R_2(P, P_0)$ will be a second order difference between P and P_0 . Applying this identity to $P = P_n^*$, and using that $P_n D^*(P_n^*) = 0$, results in the identity:

$$\Psi(P_n^*) - \Psi(P_0) = (P_n - P_0)D^*(P_n^*) + R_2(P_n^*, P_0).$$

Assuming $R_2(P_n^*, P_0) = o_P(1/\sqrt{n})$, $D^*(P_n^*)$ falls with probability tending to one in a P_0 -Donsker class, and $P_0\{D^*(P_n^*) - D^*(P_0)\}^2 \rightarrow 0$ in probability as $n \rightarrow \infty$, implies now the asymptotic efficiency of the substitution estimator $\Psi(P_n^*)$. The latter is a very weak consistency condition, so that the really crucial condition is $R_2(P_n^*, P_0) = o_P(n^{-1/2})$. To make the latter hold, it is crucial that one uses super-learning incorporating highly adaptive estimators. The Donsker class condition will hold if P_n^* is not an overfit so that its variation norm is controlled, utilizing that the class of functions with bounded variation norm is a Donsker class (van der Vaart and Wellner, 1996).

TMLE has been generalized to targeted minimum loss-based estimation (still denoted with TMLE) in which one utilizes that $\Psi(P)$ can be represented as $\Psi_1(Q(P))$ for some function Ψ_1 , where $Q(P) = \arg \min_Q PL(Q)$ can be defined as a minimizer of the risk of a loss-function $L(Q)(O)$. One notes that $D^*(P) = D_1^*(Q(P), G(P))$ for some nuisance parameter G . Given an initial estimator (Q_n^0, G_n^0) , one now defines a local least favorable submodel $Q_{n,\delta}^{0,\text{lfm}}$ so that $\left. \frac{d}{d\delta} L(Q_{n,\delta}^{0,\text{lfm}}) \right|_{\delta=0}$ spans $D^*(Q_n^0, G_n^0)$, and one updates Q_n^k by computing $\delta^k = \arg \min_{\delta} P_n L(Q_{n,\delta}^{k,\text{lfm}})$ and setting $Q_n^{k+1} = Q_{n,\delta^k}^{k,\text{lfm}}$, resulting in a TMLE (Q_n^*, G_n^0) solving $P_n D^*(Q_n^*, G_n^0) = 0$, and corresponding TMLE $\Psi_1(Q_n^*)$ of ψ_0 . To obtain a TMLE that solves additional score equations that serve a certain purpose, one could use δ of higher dimension than the target parameter, and also simultaneously update G_n^k with a submodel through G_n^k , and iterate updates of G_n^k simultaneously with the updates of Q_n^k , resulting in a TMLE (Q_n^*, G_n^*) .

In general, TMLE presents an iterative algorithm, utilizing a local parametric submodel with loss-function specific score equal to a user supplied $D(\cdot)$, that maps an initial estimator $P_n^0 \in \mathcal{M}$, or an initial estimator (Q_n^0, G_n^0) of (Q_0, G_0) , into an updated P_n^* , or (Q_n^*, G_n^*) , with improved empirical fit w.r.t. the loss-function of P_0 or (Q_0, G_0) , so that $P_n D(P_n^*) = 0$, or $P_n D(Q_n^*, G_n^*) = 0$. Due to this generality, its statistical applications are diverse and widespread, going beyond the construction of an efficient estimator of a pathwise differentiable target parameter for arbitrary semi-parametric models and pathwise differentiable target parameter mappings: collaborative targeted maximum likelihood estimation (CTMLE) for targeted estimation of the nuisance parameter in the canonical gradient (van der Laan and Rose, 2011; van der Laan and Gruber, 2010; Gruber and van der Laan, 2012; Stitelman and van der Laan, 2010; Gruber and van der Laan, 2010); cross-validated TMLE (CV-TMLE) to robustify the bias-reduction of the TMLE-step (Zheng and van der Laan, 2011; van der Laan and Rose, 2011); guaranteed improvement w.r.t. a user supplied asymptotically linear estimator (Gruber and van der Laan, 2012; Lendle et al., 2013); targeted initial estimator through empirical efficiency maximization (Rubin and van der Laan, 2008; van der Laan and Rose, 2011); double robust inference by targeting censoring/treatment mechanism (van der Laan, 2012); CV-TMLE to estimate data adaptive target parameters such as the risk of a candidate estimator and thereby develop a super-learner that uses CV-TMLE instead

of the normal cross-validated empirical risk (van der Laan and Petersen, 2012; Díaz and van der Laan, 2013, In press); higher-order TMLE in order to replace in the above proof $R_2()$ by a higher order term (Carone et al., 2014; Diaz et al., 2015).

Even though the TMLE framework has been shown to be flexible enough to handle any of the challenges we have encountered, in many cases the proposed TMLE is iterative and uses a local parametric submodel through the initial estimator that has more, and possibly many more, than d (fluctuation) parameters. This can result in small sample issues regarding convergence of the TMLE algorithm or causes finite sample instability of the estimator. It also contrasts the principle goal of TMLE as being a procedure that updates the initial estimator with *minimal* extra data fitting into a new efficient estimator. By using an over-parameterized local submodel or by using an iterative algorithm these TMLE use more fitting of the data than should be needed to achieve the desired goal.

Goal of article: The goal set out in this article is to construct a parametric submodel $\{P_{n,\epsilon}^0 : \epsilon\}$ through an initial $P_n^0 \in \mathcal{M}$ so that the above TMLE algorithm only takes one step, and the dimension of ϵ is smaller than or equal to d . The construction of this parametric submodel, a so called universal least favorable submodel, will be philosophically grounded by being in a sense the shortest path (with distance measured by information/data fitting needed) towards its goal (solving the desired score equation). We will first consider the case $d = 1$ and construct a one-dimensional parametric submodel satisfying this key property so that the TMLE is a one-step TMLE. We will generalize it to targeted minimum loss-based estimation, with all its variations in choice of loss function, and demonstrate it with various examples. Finally, we consider the general case $d > 1$, and construct a *one*-dimensional parametric submodel through P_n^0 for which the one-step TMLE solves each of the d desired score equations. Apparently, this one-dimensional path provides a “shortest” path towards its d -dimensional goal. We will show that this result extends to an infinite dimensional target parameter.

3 Intuition of TMLE: local and universal least favorable submodels

Let's consider one-dimensional target parameters (i.e., $d = 1$). A least favorable model at P is a model $\mathcal{S}^* = \{P_{\epsilon,h^*} : \epsilon\}$, dominated by P , for which $P_{\epsilon=0,h^*} = P$, and that maximizes the submodel specific Cramer-Rao lower bound for the asymptotic variance of a regular asymptotically linear estimator of $\Psi(P_{\epsilon=0})$ for submodel $\{P_{\epsilon,h} : \epsilon\}$ defined by

$$CR(h \mid P) \equiv \frac{\left(\frac{d}{d\epsilon} \Psi(P_{\epsilon,h}) \Big|_{\epsilon=0}\right)^2}{-P \frac{d^2}{d\epsilon^2} \log \frac{dP_{\epsilon,h}}{dP} \Big|_{\epsilon=0}}.$$

It maximizes $CR(h \mid P)$ over all such parametric submodels $\{P_{\epsilon,h} : \epsilon\}$ with h varying over some index set whose closure of the linear span generated the full tangent space

$T(P) \subset L_0^2(P)$ of the model at P . Given the pathwise differentiability with canonical gradient $D^*(P)$, denoting the score of $\{P_{\epsilon,h} : \epsilon\}$ at $\epsilon = 0$ with S_h , it follows that this criterion for a submodel can be represented as follows:

$$CR(h | P) = \frac{(PD^*(P)S_h)^2}{PS_h^2},$$

By the Cauchy-Schwarz inequality, it follows that this is maximized over all scores in the tangent space $T(P)$ by $S = D^*(P)$. Thus, a least favorable model can also be defined as any parametric model through P that has a score at P equal to $D^*(P)$.

By using a second order Taylor expansion of $\epsilon \rightarrow P \log dP_{\epsilon,h}/dP$ at $\epsilon = 0$ and that this equals the information PS_h^2 , it follows that, under some smoothness assumptions on the submodels, the criterion can also be represented as

$$CR(h | P) = \lim_{\epsilon \rightarrow 0} \frac{(\Psi(P_{\epsilon,h}) - \Psi(P))^2}{-2P \log dP_{\epsilon,h}/dP}.$$

This shows that $CR(h | P)$ equals the square change in the target parameter divided by the change in log-likelihood at P at an infinitesimal ϵ . Therefore, we will say that the path $\{P_{\epsilon,h^*} : \epsilon\}$ that maximizes $CR(h | P)$ follows at $\epsilon = 0$ (i.e., locally) a path of maximal change in target parameter per unit of information. To stress that the desired optimality property only applies locally, we will refer to such a submodel as a *locally* (i.e., at $\epsilon = 0$) least favorable submodel.

This latter representation of the criterion is intuitively appealing, since a sensible goal of a submodel $\{P_\epsilon : \epsilon\}$ through P is that a small fluctuation of P yields a maximal change in target parameter, making the MLE $\epsilon_n = \arg \max_\epsilon P_n \log dP_\epsilon/dP$ (as used in TMLE) for this parametric model locally all about fitting the target parameter, not wasting data for anything else.

The intuition of TMLE has always been to minimally increase the empirical fit of the initial estimator while achieving the desired bias reduction for the target parameter, measured by solving $P_n D^*(P_n^*)$ with a good estimator P_n^* of P_0 (so not worse than P_n^0). However, if P_n^0 is far away from P_0 , the MLE ϵ_n^0 will be far from local. Even though it moves in the right direction at $\epsilon \approx 0$, there is no guarantee that it follows a path of maximal change in target parameter per change in distribution once ϵ moves farther away from zero. In the end that means that the targeted maximum likelihood estimator might not have followed such a targeted path after all, and it might have taken various iterations to finally end up with a local $\epsilon_n^K \approx 0$ at which point the algorithm stops. The distribution P_n^0 might have changed much more than needed to obtain the bias reduction in the target parameter. That is, the desired bias reduction came at an unnecessary cost of data fitting so that $\Psi(P_n^*)$ will have larger finite sample variance than needed. Based on this insight, we would like to construct a TMLE that is based on a path that at each ϵ (not just at $\epsilon = 0$) follows a path of maximal change in target parameter per unit of information. We will refer to such a path as a *universal* least favorable submodel.

Definition 1 Suppose that, given a $P \in \mathcal{M}$, $U\text{lfm}(P) = \{P_\epsilon : \epsilon \in (-a, a)\} \subset \mathcal{M}$ is a parametric submodel dominated by P , such that $P_{\epsilon=0} = P$ and for each $\epsilon \in (-a, a) \subset \mathbb{R}$, we have

$$\frac{d}{d\epsilon} \log \frac{dP_\epsilon}{dP} = D^*(P_\epsilon). \quad (1)$$

Then, we say that $U\text{lfm}(P)$ is a universal least favorable submodel through P .

That is, this least favorable model is not only least favorable at $\epsilon = 0$, it is a least favorable model at each $P_\epsilon \in U\text{lfm}(P)$. This article proposes such universal least favorable submodels and corresponding targeted maximum likelihood and targeted minimum loss-based estimators. A very nice by-product of these universal least favorable submodels is that the TMLE always “converges” in one step. This reflects the above intuition of a universal least favorable submodel as a shortest path submodel in the sense that it achieves the desired bias reduction at minimal increase in empirical log-likelihood.

4 A universal least favorable submodel for targeted maximum likelihood estimation

4.1 The TMLE based on a universal least favorable submodel takes only one step

Let P_n^0 be an initial estimator of P_0 . Suppose that, given a $P \in \mathcal{M}$, we can construct a universal least favorable parametric model $U\text{lfm}(P) = \{P_\epsilon : \epsilon \in (-a, a)\} \subset \mathcal{M}$. If we use this as parametric submodel in the TMLE, then the TMLE converges in one step. That is, let

$$\epsilon_n^0 = \arg \max_{\epsilon} P_n \log \frac{dP_{n,\epsilon}^0}{dP_n^0}.$$

One can replace the maximum ϵ_n^0 by the local maximum closest to $\epsilon = 0$, which is what we recommend in case the selected universal least favorable submodel allows for multiple local maxima. Let $P_n^1 = P_{n,\epsilon_n^0}^0$. Since ϵ_n^0 is a local maximum it solves its score equation, given by $P_n D^*(P_n^1) = 0$. That is, it achieves the goal of solving the desired efficient influence curve equation in one step. Further iteration will not yield further updates: the next MLE

$$\epsilon_n^1 = \arg \max_{\epsilon} P_n \log \frac{dP_{n,\epsilon}^1}{dP_n^1} = 0.$$

Therefore, the TMLE of $\psi_0 = \Psi(P_0)$ is given by the one-step TMLE $\psi_n^* = \Psi(P_n^1)$.

In addition, we strongly suspect that a TMLE using such a least favorable model will often perform better in finite samples, certainly in situations in which the TMLE requires an iterative algorithm. In addition, it is philosophically superior by always following a path along ϵ in which the rate of square change in the parameter by unit of information is maximized at each ϵ -value.

4.2 An analytic formula for a universal least favorable submodel

This motivates us to consider if such a universal least favorable model exists and can be constructed. The answer is, yes, as our constructions below demonstrate.

In the following we use p_ϵ for the density of P_ϵ w.r.t. P , so that $p = 1$, but we will still use p (in case, one wants to use the formulas for densities w.r.t. another dominating measure). For $\epsilon \geq 0$, we recursively define

$$p_\epsilon = p \exp \left(\int_0^\epsilon D^*(P_x) dx \right), \quad (2)$$

and, for $\epsilon < 0$, we recursively define

$$p_\epsilon = p \exp \left(- \int_\epsilon^0 D^*(P_x) dx \right).$$

Theorem 1 Consider the definition of $\{P_\epsilon : \epsilon \in (-a, a)\}$ above. We have that $\{P_\epsilon : \epsilon \in (-a, a)\}$ is a set of probability distributions dominated by P , $P_{\epsilon=0} = P$, and, for each $\epsilon \in (-a, a)$, we have

$$\frac{d}{d\epsilon} \log \frac{dP_\epsilon}{dP} = D^*(P_\epsilon).$$

Proof: It follows trivially that for each ϵ , $\frac{d}{d\epsilon} \log p_\epsilon = D^*(P_\epsilon)$. It remains to verify that p_ϵ satisfies $\int p_\epsilon(o) dP(o) = 1$ (obviously, $p_\epsilon \geq 0$). Define $C(\epsilon, P) \equiv \int p_\epsilon dP$. Consider the probability density $p_{\epsilon,1} = C(\epsilon, P)^{-1} p_\epsilon$. We have that its score at ϵ is given by:

$$S(\epsilon, P) = \frac{1}{C(\epsilon, P)} \frac{d}{d\epsilon} C(\epsilon, P) + D^*(P_\epsilon).$$

We know that $P_\epsilon S(\epsilon, P) = 0$. Since $P_\epsilon D^*(P_\epsilon) = 0$, this implies that $\frac{d}{d\epsilon} C(\epsilon, P) = 0$. Thus, $C(\epsilon, P) = C(0, P) = 1$. This completes the proof. \square

Note that this recursive relation (2) allows one to recursively solve for $p_{\epsilon+d\epsilon}$, given $\{p_x : x \in [0, \epsilon]\}$, in the sense that (e.g.) for $\epsilon > 0$,

$$\frac{p_{\epsilon+d\epsilon}}{p_\epsilon} = \exp(D^*(P_\epsilon)d\epsilon) = (1 + d\epsilon D^*(P_\epsilon)).$$

This differential equation is equivalent with stating that $\frac{d}{d\epsilon} \log p_\epsilon = D^*(P_\epsilon)$. This implies a practical construction that starts with $p_{x_0=0} = p$ and recursively solves for

$$p_{x_j} = p_{x_{j-1}} (1 + (x_j - x_{j-1}) D^*(P_{x_{j-1}})), \quad j = 1, \dots, N$$

for an arbitrary fine grid $0 = x_0 < x_1 < \dots < x_N = a$. Similarly, one determines recursively

$$p_{-x_j} = p_{-x_{j+1}} (1 - (x_j - x_{j+1}) D^*(P_{-x_{j+1}})), \quad j = 1, \dots, N.$$

If the model \mathcal{M} is nonparametric, then this practical construction is a submodel of \mathcal{M} , but if the model is restricted the practical construction above might select probability distributions P_{x_j} that are not an element of \mathcal{M} , even though it has score at x_j equal to $D^*(P_{x_j})$ in the tangent space at P_{x_j} of the model \mathcal{M} . Nonetheless, this practical construction of this least favorable model can be used for any model \mathcal{M} , as long as one can extend the target parameter Ψ to be well defined on the probability distributions in this discrete approximation of the theoretical least favorable model. The TMLE will still only require one step and be asymptotically efficient for the actual model \mathcal{M} under regularity conditions. In addition, in the next subsection Theorem 2 proves that under mild regularity conditions, quite surprisingly, the theoretical formula (2) for this universal least favorable model, defined as a limit of the above practical construction when the partitioning gets finer and finer, is an actual submodel of \mathcal{M} . Another way of viewing this result is that by selecting the partitioning fine enough in the above practical construction $\{p_{x_j}, p_{-x_j} : j = 0, \dots, N\}$ this submodel will be arbitrarily close to the model \mathcal{M} . Below we will also provide an alternative to the above practical construction that does preserve the submodel property while it still approximates the theoretical formula (2).

4.3 A universal least favorable submodel in terms of a local least favorable submodel

An alternative representation of the above analytic formula (2) is given by a product integral representation. Let $d\epsilon > 0$. For $\epsilon \geq 0$, we define

$$p_{\epsilon+d\epsilon} = p \prod_{x \in (0, \epsilon]} (1 + D^*(P_x)dx),$$

and for $\epsilon < 0$, we define

$$p_{\epsilon-d\epsilon} = p \prod_{x \in [\epsilon, 0)} (1 - D^*(P_x)dx).$$

In other words, $p_{x+dx} = p_x(1 + D^*(P_x)dx)$, or, another way of thinking about this is that p_{x+dx} is obtained by constructing a least favorable model through P_x with score $D^*(P_x)$ at P_x , and evaluate it at parameter value dx , slightly away from zero. This suggests the following generalization of the universal least favorable model whose practical analogue will now still be an actual submodel of \mathcal{M} .

Let $0 = x_0 < x_1 < \dots \leq x_N = a$ be an equally spaced fine grid for the interval $[0, a]$. Let $h = x_j - x_{j-1}$ be the width of the partition elements. We will provide a construction for P_{x_j} , $j = 0, \dots, N$. This construction is expressed in terms of a mapping $P \rightarrow \{P_\delta^{\text{lfm}} : \delta \in (-a, a)\} \subset \mathcal{M}$ that maps any $P \in \mathcal{M}$ into a local least favorable submodel of \mathcal{M} through P at $\delta = 0$ and with score $D^*(P)$ at $\delta = 0$, where a is some positive number. For any estimation problem defined by \mathcal{M} and Ψ one is typically able to construct such a local least favorable submodel, so that this is hardly

an assumption. Let $P_{x=0} = P$. Let $p_{x_1} = p_{x_0, h}^{\text{lfm}}$, and, in general, let $p_{x_{j+1}} = p_{x_j, h}^{\text{lfm}}$, $j = 1, 2, \dots, N-1$. Similarly, let $-a = -x_N < -x_{N-1} < \dots < -x_1 < x_0 = 0$ be the corresponding grid for $[-a, 0]$, and we define $p_{-x_{j+1}} = p_{-x_j, -h}^{\text{lfm}}$, $j = 1, \dots, N-1$. In this manner, we have defined P_{x_j}, P_{-x_j} , $j = 0, \dots, N$, and, by construction, each of these are probability distributions in the model \mathcal{M} . The choice N defines an end value a , but one does not need to *a priori* select N . One only needs to select a small $dx = x_j - x_{j-1}$, and continue until the first local MLE is reached. This construction is all we need when using the universal least favorable submodel in practice, such as in the TMLE.

This practical construction implies a theoretical formulation by letting N converge to infinity (i.e., let the width of the partitioning converge to zero). That is, an analytic way of representing this universal least favorable submodel, given the local least favorable model parameterization $(\epsilon, P) \rightarrow p_{\epsilon, h}^{\text{lfm}}$, is given by: for $\epsilon > 0$ and $d\epsilon > 0$, we have

$$p_{\epsilon+d\epsilon} = p_{\epsilon, d\epsilon}^{\text{lfm}}.$$

This allows for the recursive solving for p_{ϵ} starting at $p_{\epsilon=0} = p$, and since $p_{\epsilon, h}^{\text{lfm}} \in \mathcal{M}$, its practical approximation will never leave the model \mathcal{M} .

Utilizing that the least favorable model $h \rightarrow p_{\epsilon, h}^{\text{lfm}}$ is continuously twice differentiable with a score $D^*(P_{\epsilon})$ at $h = 0$, we obtain a second order Taylor expansion

$$p_{\epsilon, d\epsilon}^{\text{lfm}} = p_{\epsilon} + \left. \frac{d}{dh} p_{\epsilon, h}^{\text{lfm}} \right|_{h=0} d\epsilon + O((d\epsilon)^2) = p_{\epsilon}(1 + d\epsilon D^*(P_{\epsilon})) + O((d\epsilon)^2),$$

so that we obtain

$$p_{\epsilon+d\epsilon} = p_{\epsilon}(1 + d\epsilon D^*(P_{\epsilon})) + O((d\epsilon)^2).$$

This implies:

$$p_{\epsilon} = p \exp \left(\int_0^{\epsilon} D^*(P_x) dx \right).$$

Thus, we obtained the exact same representation (2) as above. This proves that, under mild regularity conditions, this analytic representation (2) is a submodel of \mathcal{M} after all, but, when using its practical implementation and approximation, one should use an actual local least favorable submodel in order to guarantee that one stays in the model. We formalize this result in the following theorem.

Theorem 2 *Let \mathcal{O} be a maximal support so that the support of a $P \in \mathcal{M}$ is a subset of \mathcal{O} . Suppose there exists a mapping $P \rightarrow \{P_{\delta}^{\text{lfm}} : \delta \in (-a, a)\} \subset \mathcal{M}$ that maps any $P \in \mathcal{M}$ into a local least favorable submodel of \mathcal{M} through P at $\delta = 0$ and with score $D^*(P)$ at $\delta = 0$, where a is some positive number independent of P . In addition, assume the following type of second order Taylor expansion:*

$$p_{\epsilon, d\epsilon}^{\text{lfm}} = p_{\epsilon} + \left. \frac{d}{dh} p_{\epsilon, h}^{\text{lfm}} \right|_{h=0} d\epsilon + R_2(p_{\epsilon}, d\epsilon),$$

where

$$\sup_{\epsilon} \sup_{o \in \mathcal{O}} |R_2(p_{\epsilon}, d\epsilon)(o)| = O((d\epsilon)^2).$$

We also assume that $\sup_{\epsilon} \sup_{o \in \mathcal{O}} |D^*(P_{\epsilon})p_{\epsilon}|(o) < \infty$.

Then, the universal least favorable $\{p_{\epsilon} : \epsilon\}$ defined by (2) is an actual submodel of \mathcal{M} . Its definition corresponds with $p_{\epsilon+d\epsilon} = p_{\epsilon, d\epsilon}^{\text{lfm}}$ whose corresponding practical approximation will still be a submodel.

We refer to the Appendix for an application of the universal least favorable submodel and a corresponding one-step TMLE for high dimensional parametric models.

5 Universal least favorable model for targeted minimum loss-based estimation

5.1 A universal least favorable submodel w.r.t. specific loss-function

Let's now generalize this construction of a universal least favorable w.r.t. log-likelihood loss to general loss-functions so that the resulting universal least favorable submodels can be used in the more general targeted minimum loss based estimation methodology. We now assume that $\Psi(P) = \Psi_1(Q(P))$ for some parameter $Q : \mathcal{M} \rightarrow Q(\mathcal{M})$ defined on the model and real valued function Ψ_1 . Here $Q(\mathcal{M}) = \{Q(P) : P \in \mathcal{M}\}$ denotes the parameter space of this parameter Q . Let $L(Q)(O)$ be a loss-function for $Q(P)$ in the sense that $Q(P) = \arg \min_{Q \in Q(\mathcal{M})} PL(Q)$. With slight abuse of notation, let $D^*(P) = D^*(Q(P), G(P))$ be the canonical gradient of Ψ at P , where $G : \mathcal{M} \rightarrow G(\mathcal{M})$ is some nuisance parameter. We consider the case that the efficient influence curve is in the tangent space of Q , so that a least favorable submodel does not need to fluctuate G : otherwise, just include G in the definition of Q . Given, (Q, G) , let $\{Q_{\epsilon}^{\text{lfm}} : \epsilon \in (-a, a)\} \subset Q(\mathcal{M})$ be a local least favorable model w.r.t. loss function $L(Q)$ at $\epsilon = 0$ so that

$$\left. \frac{d}{d\epsilon} L(Q_{\epsilon}^{\text{lfm}}) \right|_{\epsilon=0} = D^*(Q, G).$$

The dependence of this submodel on G is suppressed in this notation.

Let $0 = x_0 < x_1 < \dots < x_N = a$ be an equally spaced fine grid for the interval $[0, a]$. Let $h = x_j - x_{j-1}$ be the width of the partition elements. We present a construction for Q_{x_j} , $j = 0, \dots, N$. Let $Q_{x=0} = Q$. Let $Q_{x_1} = Q_{x_0, h}^{\text{lfm}}$, and, in general, let $Q_{x_{j+1}} = Q_{x_j, h}^{\text{lfm}}$, $j = 1, 2, \dots, N-1$. Similarly, let $-a = -x_N < -x_{N-1} < \dots < -x_1 < x_0 = 0$ be the corresponding grid for $[-a, 0]$, and we define $Q_{-x_{j+1}} = Q_{-x_j, -h}^{\text{lfm}}$, $j = 1, \dots, N-1$. In this manner, we have defined Q_{x_j}, Q_{-x_j} , $j = 0, \dots, N$, and, by construction, each of these are an element of the parameter

space $Q(\mathcal{M})$. This construction is all we need when using this submodel in practice, such as in the TMLE.

An analytic way of representing this loss-function specific universal least favorable submodel for $\epsilon \geq 0$ (and similarly for $\epsilon < 0$) is given by: for $\epsilon > 0$, $d\epsilon > 0$,

$$Q_{\epsilon+d\epsilon} = Q_{\epsilon,d\epsilon}^{\text{lfm}}, \quad (3)$$

allowing for the recursive solving for Q_ϵ starting at $Q_{\epsilon=0} = Q$, and since $Q_{\epsilon,h}^{\text{lfm}} \in Q(\mathcal{M})$, its practical approximation never leaves the parameter space $Q(\mathcal{M})$ for Q .

Let's now derive a corresponding integral equation. Assume that for some $\dot{L}(Q)(O)$, we have

$$\left. \frac{d}{dh} L(Q_{\epsilon,h}^{\text{lfm}}) \right|_{h=0} = \dot{L}(Q_\epsilon) \left. \frac{d}{dh} Q_{\epsilon,h}^{\text{lfm}} \right|_{h=0}.$$

Then, by the local property of a least favorable submodel,

$$\left. \frac{d}{dh} Q_{\epsilon,h}^{\text{lfm}} \right|_{h=0} = \frac{D^*(Q_\epsilon, G)}{\dot{L}(Q_\epsilon)}.$$

Utilizing that the local least favorable model $h \rightarrow Q_{\epsilon,h}^{\text{lfm}}$ is twice continuously differentiable with derivative $D^*(Q_\epsilon, G)/\dot{L}(Q_\epsilon)$ at $h = 0$, we obtain the following second order Taylor expansion:

$$\begin{aligned} Q_{\epsilon,d\epsilon}^{\text{lfm}} &= Q_\epsilon + \left. \frac{d}{dh} Q_{\epsilon,h}^{\text{lfm}} \right|_{h=0} d\epsilon + O((d\epsilon)^2) \\ &= Q_\epsilon + \frac{D^*(Q_\epsilon, G)}{\dot{L}(Q_\epsilon)} d\epsilon + O((d\epsilon)^2). \end{aligned}$$

Note that Q_ϵ can also be represented as $Q_{\epsilon,0}^{\text{lfm}}$. This implies the following recursive analytic definition of the universal least favorable model through Q :

$$Q_\epsilon = Q + \int_0^\epsilon \frac{D^*(Q_x, G)}{\dot{L}(Q_x)} dx. \quad (4)$$

Similarly, for $\epsilon < 0$, we obtain

$$Q_\epsilon = Q - \int_\epsilon^0 \frac{D^*(Q_x, G)}{\dot{L}(Q_x)} dx.$$

As with the log-likelihood loss (and thus $Q(P) = P$), this shows that, under regularity conditions, this analytic representation for Q_ϵ is an element in $Q(\mathcal{M})$, although using it in a practical construction (in which integrals are replaced by sums) might easily leave the model space $Q(\mathcal{M})$, so that our above practical construction in terms of the local least favorable model and discrete grid represents the desired practical implementation of this universal least favorable submodel. The following theorem formalizes this result stating that the analytic formulation (4) is indeed a universal least favorable submodel.

Theorem 3 Given, any (Q, G) compatible with model \mathcal{M} , let $\{Q_\delta^{\text{lfm}} : \delta \in (-a, a)\} \subset Q(\mathcal{M})$ be a local least favorable model w.r.t. loss function $L(Q)$ at $\delta = 0$ so that

$$\left. \frac{d}{d\delta} L(Q_\delta^{\text{lfm}}) \right|_{\delta=0} = D^*(Q, G).$$

Assume that for some $\dot{L}(Q)(O)$, we have

$$\left. \frac{d}{d\epsilon} L(Q_\epsilon^{\text{lfm}}) \right|_{\epsilon=0} = \dot{L}(Q) \left. \frac{d}{d\epsilon} Q_\epsilon^{\text{lfm}} \right|_{\epsilon=0}.$$

Consider the corresponding model $\{Q_\epsilon : \epsilon\}$ defined by (4). It goes through Q at $\epsilon = 0$, and, it satisfies that for all ϵ

$$\frac{d}{d\epsilon} L(Q_\epsilon) = D^*(Q_\epsilon, G). \quad (5)$$

In addition, suppose that the $a > 0$ in the local least-favorable submodel above can be chosen to be independent of the choice $(Q, G) \in \{Q_\epsilon, G_\epsilon : \epsilon\}$, and assume the following second order Taylor expansion:

$$\begin{aligned} Q_{\epsilon, d\epsilon}^{\text{lfm}} &= Q_\epsilon + \left. \frac{d}{dh} Q_{\epsilon, h}^{\text{lfm}} \right|_{h=0} d\epsilon + R_2(Q_\epsilon, G, d\epsilon) \\ &= Q_\epsilon + \frac{D^*(Q_\epsilon, G)}{\dot{L}(Q_\epsilon)} d\epsilon + R_2(Q_\epsilon, G, d\epsilon), \end{aligned}$$

where

$$\sup_{\epsilon} \sup_{o \in \mathcal{O}} |R_2(Q_\epsilon, G, d\epsilon)(o)| = O((d\epsilon)^2).$$

We also assume that $\sup_{\epsilon} \sup_{o \in \mathcal{O}} \left| \frac{D^*(Q_\epsilon, G)}{\dot{L}(Q_\epsilon)}(o) \right| < \infty$.

Then, we also have $\{Q_\epsilon : \epsilon\} \subset Q(\mathcal{M})$.

Proof: Let $\epsilon > 0$. We have

$$\begin{aligned} \frac{d}{d\epsilon} L \left(Q + \int_0^\epsilon \frac{D^*(Q_x, G)}{\dot{L}(Q_x)} dx \right) &= \dot{L}(Q_\epsilon) \frac{d}{d\epsilon} Q_\epsilon \\ &= \dot{L}(Q_\epsilon) \frac{D^*(Q_\epsilon, G)}{\dot{L}(Q_\epsilon)} \\ &= D^*(Q_\epsilon, G). \end{aligned}$$

This completes the proof of (5). The submodel statement was already shown above, but we now provided formal sufficient conditions. \square

5.2 Example demonstrating that analytic formula (4) for universal least favorable submodel is indeed a submodel

Suppose $O = (W, A, Y) \sim P_0$, $A \in \{0, 1\}$ binary, $Y \in \{0, 1\}$ also binary, and let the statistical model \mathcal{M} be the nonparametric model or any model that only restricts the tangent space of the conditional distribution of A , given W . Let $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ be defined by $\Psi(P) = E_P E_P(Y \mid A = 1, W)$. The efficient influence curve $D^*(P)(O) = A/\bar{g}(W)(Y - \bar{Q}(W)) + \bar{Q}(W) - \Psi(P)$, where $\bar{g}(W) = P(A = 1 \mid W)$ and $\bar{Q}(W) = E_P(Y \mid A = 1, W)$. We note that $\Psi(P) = \Psi_1(Q) = Q_W \bar{Q}$, where $Q = (Q_W, \bar{Q})$, and Q_W is the probability distribution of W under P . We can decompose $D^*(P) = D_1^*(\bar{Q}, \bar{g}) + D_0^*(Q)$, where $D_1^*(\bar{Q}, \bar{g}) = A/\bar{g}(Y - \bar{Q}(W))$ is a score of the conditional distribution of Y , given A, W , while $D_0^*(Q)$ is a score of the marginal distribution of W . Since we estimate $Q_{W,0}$ with the empirical probability distribution of W_1, \dots, W_n , there is no need to construct a submodel through Q_W , so that we focus on constructing a submodel through \bar{Q} only.

A valid loss function for \bar{Q} is given by

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

Consider the local least favorable submodel through \bar{Q} :

$$\text{Logit} \bar{Q}_\epsilon^{\text{lfm}} = \text{Logit} \bar{Q} - \epsilon H(\bar{g}),$$

where $H(\bar{g})(A, W) = A/\bar{g}(W)$. This is indeed a local least favorable submodel for \bar{Q} since

$$\left. \frac{d}{d\epsilon} L(\bar{Q}_\epsilon^{\text{lfm}}) \right|_{\epsilon=0} = D_1^*(\bar{Q}, \bar{g}).$$

Let's now compute the corresponding theoretical universal least favorable submodel (4). We have

$$\frac{d}{d\epsilon} L(\bar{Q}_\epsilon) = \frac{d}{d\epsilon} Q_\epsilon \left\{ -I(A = 1) \frac{Y - \bar{Q}_\epsilon}{\bar{Q}_\epsilon(1 - \bar{Q}_\epsilon)} \right\}.$$

Thus,

$$\dot{L}(Q_\epsilon) = -I(A = 1) \frac{Y - \bar{Q}_\epsilon}{\bar{Q}_\epsilon(1 - \bar{Q}_\epsilon)}.$$

Thus, the universal least favorable submodel (4) through Q is given by:

$$\bar{Q}_\epsilon = \bar{Q} - H(\bar{g}) \int_0^\epsilon \bar{Q}_x(1 - \bar{Q}_x) dx.$$

This integral equation shows that

$$\frac{\frac{d}{d\epsilon} \bar{Q}_\epsilon}{\bar{Q}_\epsilon(1 - \bar{Q}_\epsilon)} = -H(\bar{g}).$$

This has as solution $\bar{Q}_\epsilon = Q_\epsilon^{\text{lfm}}$, and since there is only one solution, this proves that the universal least favorable submodel $\bar{Q}_\epsilon = Q_\epsilon^{\text{lfm}}$. Indeed, it follows directly that for all ϵ

$$\frac{d}{d\epsilon} L(\bar{Q}_\epsilon^{\text{lfm}}) = D_1^*(\bar{Q}_\epsilon^{\text{lfm}}, \bar{g}),$$

showing that our local least favorable submodel is already a universal least favorable submodel. Indeed, the TMLE using Q_ϵ^{lfm} requires only one step. In particular, as predicted by our theory, this demonstrates that the analytic formula (4) respects the constraints that $\bar{Q} \in (0, 1)$, even though that is not immediately obvious from its analytic integral or differential representation.

We refer to supplementary material for the construction of a universal least favorable submodels to general loss functions that are allowed to depend on an unknown nuisance parameter, and corresponding example from the causal inference literature. These examples also demonstrate that in examples for which the TMLE based on the local least favorable model already converged in one step, the least favorable submodel is actually already a universal least favorable submodel.

6 Example: One-step TMLE of average treatment effect among the treated

Let $O = (W, A, Y) \sim P_0$ and let \mathcal{M} be a nonparametric statistical model. Let $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ be defined by $\Psi(P) = E_P(E_P(Y \mid A = 1, W) - E_P(Y \mid A = 0, W) \mid A = 1)$. The efficient influence curve of Ψ at P is given by (Zheng et al., 2013):

$$D^*(P)(O) = H_1(g, q)(A, W)(Y - \bar{Q}(A, W)) + \frac{A}{q} \{\bar{Q}(1, W) - \bar{Q}(0, W) - \Psi(P)\},$$

where $g(a \mid W) = P(A = a \mid W)$, $\bar{Q}(a, W) = E_P(Y \mid A = a, W)$, $q = P(A = 1)$, and

$$H_1(g, q)(A, W) = \frac{A}{q} - \frac{(1 - A)g(1 \mid W)}{qg(0 \mid W)}.$$

We note that

$$\Psi(P) = \Psi_1(Q_W, \bar{Q}, g, q) = \int \{\bar{Q}(1, w) - \bar{Q}(0, w)\} \frac{g(1 \mid w)}{q} dQ_W(w),$$

where Q_W is the probability distribution of W under P . So, if we define $Q = (Q_W, \bar{Q}, g, q)$, then $\Psi(P) = \Psi_1(Q)$. For notational convenience, we will use $\Psi(P)$ and $\Psi(Q)$ interchangeably. Since we can estimate Q_W and q with their empirical probability distributions, we are only interested in a universal least favorable submodel for (\bar{Q}, g) . We can orthogonally decompose $D^*(P) = D_1^*(P) + D_2^*(P) + D_3^*(P)$

in $L_0^2(P)$ into scores of \bar{Q} , g , and Q_W , respectively, where

$$\begin{aligned} D_1^*(P) &= H_1(g, q)(A, W)(Y - \bar{Q}(A, W)) \\ D_2^*(P) &= H_2(Q)(W)(A - g(1 | W)) \\ D_3^*(P) &= \frac{g(1 | W)}{q} \{ \bar{Q}(1, W) - \bar{Q}(0, W) - \Psi(Q) \}, \end{aligned}$$

and

$$H_2(Q)(W) = \frac{\bar{Q}(1, W) - \bar{Q}(0, W) - \Psi(Q)}{q}.$$

Thus the component of the efficient influence curve corresponding with (\bar{Q}, g) is given by $D_1^*(Q) + D_2^*(Q)$.

We consider the following loss-functions and local least favorable submodels for \bar{Q} and g (Zheng et al., 2013):

$$\begin{aligned} L_1(\bar{Q})(O) &= -\{Y \log \bar{Q}(A, W) + (1 - Y) \log(1 - \bar{Q}(A, W))\} \\ \text{Logit} \bar{Q}_\epsilon^{\text{lfm}} &= \text{Logit} \bar{Q} - \epsilon H_1(g, q) \\ L_2(g)(O) &= -\{A \log g(1 | W) + (1 - A) \log g(0 | W)\} \\ \text{Logit} \bar{g}_\epsilon^{\text{lfm}} &= \text{Logit} \bar{g} - \epsilon H_2(Q). \end{aligned}$$

We now define the sum loss function $L(\bar{Q}, g) = L_1(\bar{Q}) + L_2(g)$ and local least favorable submodel $\{\bar{Q}_\epsilon^{\text{lfm}}, g_\epsilon^{\text{lfm}} : \epsilon\}$ through (\bar{Q}, g) at $\epsilon = 0$ satisfying

$$\left. \frac{d}{d\epsilon} L(\bar{Q}_\epsilon^{\text{lfm}}, g_\epsilon^{\text{lfm}}) \right|_{\epsilon=0} = D_1^*(Q) + D_2^*(Q).$$

Thus, we can conclude that this defines indeed a local least favorable submodel for (\bar{Q}, g) .

The universal least favorable submodel (3) is now defined by the following recursive definition: for $\epsilon \geq 0$ and $d\epsilon > 0$,

$$\begin{aligned} \text{Logit} \bar{Q}_{\epsilon+d\epsilon} &= \text{Logit} \bar{Q}_{\epsilon, d\epsilon}^{\text{lfm}} \\ &= \text{Logit} \bar{Q}_\epsilon - d\epsilon H_1(g_\epsilon, q) \\ \text{Logit} \bar{g}_{\epsilon+d\epsilon} &= \text{Logit} \bar{g}_{\epsilon, d\epsilon}^{\text{lfm}} \\ &= \text{Logit} \bar{g}_\epsilon - d\epsilon H_2(Q_W, \bar{Q}_\epsilon, q). \end{aligned}$$

Similarly, we have a recursive relation for $\epsilon < 0$, but since all these formulas are just symmetric versions of the $\epsilon > 0$ case, we will focus on $\epsilon > 0$. This expresses the next $(\bar{Q}_{\epsilon+d\epsilon}, g_{\epsilon+d\epsilon})$ in terms of previously calculated $(\bar{Q}_x, g_x : x \leq \epsilon)$, thereby fully defining this universal least favorable submodel. This recursive definition corresponds with the following integral representation of this universal least favorable submodel:

$$\begin{aligned} \text{Logit} \bar{Q}_\epsilon &= \text{Logit} \bar{Q} - \int_0^\epsilon H_1(g_x, q) dx \\ \text{Logit} \bar{g}_\epsilon &= \text{Logit} \bar{g} - \int_0^\epsilon H_2(Q_W, \bar{Q}_x, q) dx. \end{aligned}$$

Let's now explicitly verify that this indeed satisfies the key property of a universal least favorable submodel. Clearly, it is a submodel and it contains (Q, g) at $\epsilon = 0$. The score of \bar{Q}_ϵ at ϵ is given by $H_1(g_\epsilon, q)(Y - \bar{Q}_\epsilon)$ and the score of g_ϵ at ϵ is given by $H_2(Q_W, \bar{Q}_\epsilon, q)(A - \bar{g}_\epsilon(W))$, so that

$$\begin{aligned} \frac{d}{d\epsilon} L(\bar{Q}_\epsilon, g_\epsilon) &= H_1(g_\epsilon, q)(Y - \bar{Q}_\epsilon) + H_2(Q_W, \bar{Q}_\epsilon, q)(A - \bar{g}_\epsilon(W)) \\ &= D_1^*(Q_W, \bar{Q}_\epsilon, g_\epsilon, q) + D_2^*(Q_W, \bar{Q}_\epsilon, g_\epsilon, q), \end{aligned}$$

explicitly proving that indeed this is a universal least favorable model for (\bar{Q}, g) .

In our previous work on the TMLE for the average treatment effect among the treated we implemented the TMLE based on the local least favorable submodel $\{\bar{Q}_{\epsilon_1}^{\text{lfm}}, \bar{g}_{\epsilon_2}^{\text{lfm}} : \epsilon_1, \epsilon_2\}$, using a separate ϵ_1 and ϵ_2 for \bar{Q} and \bar{g} . This TMLE can also be implemented using a single ϵ by regressing a dependent variable vector (Y, A) on a stacked design matrix consisting of an offset and covariate H , the vector $(H_1(g, q)(A, W), H_2(Q)(W))$. This TMLE require several iterations until convergence, whether it is implemented using using a single ϵ or separate (ϵ_1, ϵ_2) .

The TMLE based on the universal least favorable submodel above is implemented as follows, given an initial estimator (\bar{Q}, g) . One first determines the sign of the derivative at $h = 0$ of $P_n L(\bar{Q}_h, g_h)$. Suppose that the derivative is negative so that it decreases for $h > 0$. Then, one keeps iteratively calculating $(\bar{Q}_{\epsilon+d\epsilon}, g_{\epsilon+d\epsilon})$ for small $d\epsilon > 0$, given $(\bar{Q}_x, g_x : x \leq \epsilon)$, till $P_n L(\bar{Q}_{\epsilon+d\epsilon}, g_{\epsilon+d\epsilon}) \geq P_n L(\bar{Q}_\epsilon, g_\epsilon)$, at which point the desired local maximum likelihood ϵ_n is attained. The TMLE of (\bar{Q}_0, g_0) is now given by $\bar{Q}_{\epsilon_n}, g_{\epsilon_n}$, which solves $P_n\{D_1^*(Q_{\epsilon_n}) + D_2^*(Q_{\epsilon_n})\} = 0$, where $Q_{\epsilon_n} = (Q_{W,n}, \bar{Q}_{\epsilon_n}, g_{\epsilon_n}, q_n)$, and $Q_{W,n}, q_n$ are the empirical counterparts of $Q_{W,0}, q_0$. Since, we also have $P_n D_3^*(Q_{\epsilon_n}) = 0$, it follows that $P_n D^*(Q_{\epsilon_n}) = 0$. The (one-step) TMLE of $\Psi(Q_0)$ is given by the corresponding plug-in estimator $\Psi(Q_{\epsilon_n})$.

7 Simulation Studies for the average treatment effect among the treated

The iterative TMLE for estimating the average treatment effect among the treated (ATT) parameter returns to the data several times to make a sequence of local moves that updates the estimate of $\bar{Q}_n(A, W)$ and $\bar{g}_n(A, W)$ at each iteration. In contrast, the one-step TMLE using the universal least favorable sub-model fits the data once, where the MLE step requires a series of micro updates within a much smaller local neighborhood defined by a tuning parameter step size, $d\epsilon$. When there is sufficient information in the data for estimating the target parameter these two approaches can be expected to have comparable performance. When there is sparsity in the data theory suggests the one-step TMLE will be more stable, having lower variance than the iterative TMLE.

Two simulation studies demonstrate these properties. The iterative TMLE was implemented using a single ϵ , the closest analog to the one-step TMLE. $d\epsilon$ was set

to 0.001 for the one-step TMLE. Source code for the estimators and the simulation studies is available as supplementary materials. The parameter of interest is defined by the mapping $\Psi_1(Q) = \int \{\bar{Q}(1, w) - \bar{Q}(0, w)\} \frac{g(1|w)}{q} dQ_w(w)$. Each TMLE targets initial estimates $\bar{Q}_n^0(A, W)$ and $g_n^0(W)$ towards the parameter of interest. The parameter estimate is evaluated by plugging the updated estimates, $\bar{Q}_n^*(A, W), g_n^*(W)$ into the mapping, with the integral approximated by taking the empirical mean over all observations in the data, $\psi_n = \frac{1}{n} \sum_{i=1}^n \{\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)\} \frac{g_n^*(1|W_i)}{q}$.

Simulation Study I: For this study 1000 datasets were generated at two sample sizes, $n = 100$ and $n = 1000$. Two normally distributed covariates and one binary covariate were generated as $W_1 \sim N(0, 1)$, $W_2 \sim N(0, 1)$, $W_3 \sim \text{Bern}(0.5)$. All covariates are independent. Treatment assignment probabilities are given by $P(A = 1 | W) = \text{expit}(-0.4 - 0.2W_1 - 0.4W_2 + 0.3W_3)$. A binary outcome, Y was generated by setting $P(Y = 1 | A, W) = \text{expit}(-1.2 - 1.2A - 0.1W_1 - 0.2W_2 - 0.1W_3)$. The true value of the ATT parameter is $\psi_0 = -0.1490$. There are no theoretical positivity violations (treatment assignment probabilities were typically between 0.07 and 0.87), but at the smaller sample size there is less information in the data for estimating g within some strata of W . This suggests that some of the generated data sets will prove more challenging to the iterative TMLE than to the one-step TMLE. Estimates were obtained using correct and misspecified logistic regression models for the initial estimates of Q and g . Q_{cor} was estimated using a logistic regression of Y on A, W_1, W_2, W_3 . Q_{mis} was estimated using a logistic regression of Y on A, W_1 .

g_{cor} was estimated using a logistic regression of A on W_1, W_2, W_3 , and g_{mis} was estimated using a logistic regression of A on W_1 . Bias, variance, mean squared error (MSE), and relative efficiency ($RE = \text{MSE}_{\text{one-step}} / \text{MSE}_{\text{iter}}$) are shown in Table 1. $RE < 1$ indicates the one-step TMLE has better finite sample efficiency than the iterative TMLE.

Results: The one-step and iterative TMLEs exhibit similar performance when $n = 1000$, with $RE = 1$. When $n = 100$ the iterative TMLE failed to converge for 24 of the 1000 datasets. The performance of the two TMLEs on the remaining 976 datasets was quite similar. However, the fact that the bias, variance, and MSE of the one-step TMLE are larger when evaluated over all 1000 datasets tells us that the 24 omitted datasets where the iterative TMLE failed were among the most challenging. One way to repair the performance of the iterative TMLE is to bound predicted outcome probabilities away from 0 and 1. We re-analyzed the same 1000 datasets enforcing bounds on \bar{Q}_n of $(10^{-9}, 1 - 10^{-9})$ for both estimators. This minimal bounding prevents the iterative TMLE from failing, and should not introduce truncation bias. Bounding \bar{Q}_n allowed the iterative TMLE to produce a result for all analyses. Enforcing bounds had no effect on estimates produced by the one-step TMLE. This confirms that the strategy of taking many small steps within a local neighborhood whose boundaries shift minutely with each iteration helps avoid

extremes. Although the iterative TMLE no longer failed when \bar{Q}_n was bounded, it had higher variance and MSE than the one-step TMLE. Efficiency gains of the

Table 1: Simulation Study I. Bias, variance, mean squared error (MSE) and relative efficiency (RE) of the one-step TMLE and iterative TMLE over 1000 Monte Carlo simulations ($n = 1000$ and $n = 100$). Results when $n = 100$ are shown with and without omitting 24 challenging runs from the analysis, and when \bar{Q}_n is bounded away from 0 and 1 for both TMLEs.*

	Bias		Variance		MSE		RE
	one-step	iterative	one-step	iterative	one-step	iterative	
n = 1000							
Q correct							
g_{cor}	-0.00042	-0.00042	0.00059	0.00059	0.00059	0.00059	1.00
g_{mis}	-0.00050	-0.00050	0.00057	0.00057	0.00057	0.00057	1.00
Q misspecified							
g_{cor}	-0.00035	-0.00035	0.00059	0.00059	0.00059	0.00059	1.00
g_{mis}	0.01210	0.01210	0.00049	0.00048	0.00063	0.00063	1.00
n = 100, all runs							
Q correct							
g_{cor}	0.00049		0.00694		0.00693		
g_{mis}	-0.00215		0.00635		0.00635		
Q misspecified							
g_{cor}	0.00113		0.00685		0.00684		
g_{mis}	0.01226		0.00528		0.00543		
n = 100, (24 runs omitted)							
Q correct							
g_{cor}	0.00296	0.00295	0.00679	0.00678	0.00679	0.00679	1.00
g_{mis}	0.00023	0.00023	0.00621	0.00621	0.00621	0.00620	1.00
Q misspecified							
g_{cor}	0.00357	0.00363	0.00671	0.00669	0.00671	0.00670	1.00
g_{mis}	0.01474	0.01473	0.00509	0.00509	0.00530	0.00530	1.00
n = 100, Q bounded							
Q correct							
g_{cor}	0.00049	-0.00182	0.00694	0.00781	0.00693	0.00781	0.89
g_{mis}	-0.00215	-0.00168	0.00635	0.01033	0.00635	0.01033	0.62
Q misspecified							
g_{cor}	0.00113	-0.00052	0.00685	0.00738	0.00684	0.00738	0.93
g_{mis}	0.01226	0.01031	0.00528	0.00592	0.00543	0.00602	0.90

*bounding \bar{Q}_n had no effect on estimates produced when $n = 1000$.

one-step TMLE were between 7 and 28 percent.

Simulation Study II: This study more closely examines estimator performance when there is sparsity in the data. Sparsity was introduced by overfitting the initial \bar{Q}_n^0 , leaving little signal for the targeting step. Theory suggests the one-step TMLE will be a more stable estimator than the iterative TMLE under these challenging conditions. To explore the impact of overfitting the data on the iterative and one-step TMLEs we constructed a nested sequence of correct logistic regression outcome models. Covariates W_1, W_2, W_3 were generated as above. Eight additional independent and identically distributed covariates W_4, \dots, W_{12} were drawn from a normal distribution with mean 0 and standard deviation 1. None of the additional covariates were causally related to Y or A . The binary treatment indicator, A was generated in the same way as in study I. The outcome was generated by setting $P(Y = 1 | A, W) = \text{expit}(-1.2 - 1.2A - 0.1W_1 - 0.2W_2 - 0.1W_3)$. The smallest correct model, Q_{c1} , regresses Y on A, W_1, W_2, W_3 . Subsequent models were constructed by adding a single covariate to the model. The ten nested models were defined as

$$\begin{aligned} Q_{c1} : E[Y|A, W] &= \text{expit}(A + W_1 + W_2 + W_3), \\ Q_{c2} : E[Y|A, W] &= \text{expit}(A + W_1 + W_2 + W_3 + W_4), \\ Q_{c3} : E[Y|A, W] &= \text{expit}(A + W_1 + W_2 + W_3 + W_4 + W_5), \\ Q_{c4} : E[Y|A, W] &= \text{expit}(A + W_1 + W_2 + W_3 + W_4 + W_5 + W_6), \\ Q_{c5} : E[Y|A, W] &= \text{expit}(A + W_1 + W_2 + W_3 + W_4 + W_5 + W_6 + W_7), \\ Q_{c6} : E[Y|A, W] &= \text{expit}(A + W_1 + W_2 + W_3 + W_4 + W_5 + W_6 + W_7 + W_8), \\ Q_{c7} : E[Y|A, W] &= \text{expit}(A + W_1 + W_2 + W_3 + W_4 + W_5 + W_6 + W_7 + W_8 + W_9), \\ Q_{c8} : E[Y|A, W] &= \text{expit}(A + W_1 + W_2 + W_3 + W_4 + W_5 + W_6 + W_7 + W_8 + W_9 + W_{10}), \\ Q_{c9} : E[Y|A, W] &= \text{expit}(A + W_1 + W_2 + W_3 + W_4 + W_5 + W_6 + W_7 + W_8 + W_9 + W_{10} \\ &\quad + W_{11}), \\ Q_{c10} : E[Y|A, W] &= \text{expit}(A + W_1 + W_2 + W_3 + W_4 + W_5 + W_6 + W_7 + W_8 + W_9 + W_{10} \\ &\quad + W_{11} + W_{12}). \end{aligned}$$

Each of these regression models is correct, but as the model grows larger and larger the model fitting procedure begins to respond to random variation in the outcome. This problem is more acute at smaller sample sizes.

Estimates were obtained from 1000 datasets ($n = 100$), with g modeled correctly as a regression of A on W_1, W_2, W_3 . Bias, variance, MSE, and RE are reported in Table 2. The iterative TMLE failed on a large number of datasets. On the less challenging datasets where it did converge, performance of the iterative and one-step TMLEs was quite similar. When bounds on \bar{Q}_n were enforced at $(10^{-9}, 1 - 10^{-9})$, the performance of the one-step TMLE was unchanged, while the iterative TMLE was repaired. The iterative TMLE had larger bias, variance, and MSE than the one-step TMLE, which was up to four times more efficient than the iterative TMLE. These results are plotted in Fig. 1, along with estimates obtained when the parameter was evaluated based on each initial non-targeted outcome regression fit. The behavior

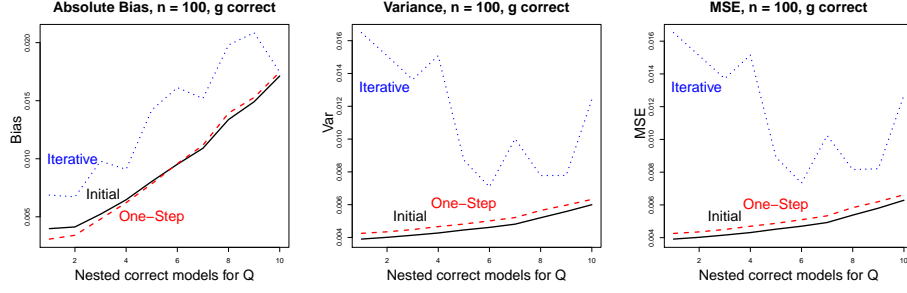


Figure 1: Simulation study II. Bias, Var, MSE for iterative TMLE, one-step TMLE, and the non-targeted Initial estimator as overfit increases, $Q_{c1} - Q_{c10}$.

of the iterative TMLE is erratic, while that of the non-targeted estimator and the one-step TMLE are quite stable.

8 Universal canonical one-dimensional submodel that targets a multidimensional target parameter

Let $\Psi : \mathcal{M} \rightarrow H$ be a Hilbert-space valued pathwise differentiable target parameter. Typically, we simply have $H = \mathbb{R}^d$ endowed with the standard inner product $\langle x, y \rangle = \sum_{j=1}^d x_j y_j$. However, we also allow that $\Psi(P)$ is a function $t \rightarrow \Psi(P)(t)$ from $\tau \subset \mathbb{R}$ to \mathbb{R} in a Hilbert space $L^2(\Lambda)$ endowed with inner product $\langle h_1, h_2 \rangle = \int h_1(t) h_2(t) d\Lambda(t)$, where Λ is a user supplied positive measure with $\int d\Lambda(t) < \infty$. For notational convenience, we will often denote the inner product $\langle h_1, h_2 \rangle$ with $h_1^\top h_2$, analogue to the typical notation for the inner product in \mathbb{R}^d . Let $\|h\| = \sqrt{\langle h, h \rangle}$ be the Hilbert space norm, which would be the standard Euclidean norm in the case that $H = \mathbb{R}^d$. Let $D^*(P)$ be the canonical gradient. If $H = \mathbb{R}^d$, then this is a d -dimensional canonical gradient $D^*(P) = (D_j^*(P) : j = 1, \dots, d)$, but in general $D^*(P) = (D_t^*(P) : t \in \tau)$. Let $L(p) = -\log p$, where $p = dP/d\mu$ is a density of $P \ll \mu$ w.r.t. some dominating measure μ . In this section we will construct a one-dimensional submodel $\{P_\epsilon : \epsilon \geq 0\}$ through P at $\epsilon = 0$ so that, for any $\epsilon \geq 0$,

$$\frac{d}{d\epsilon} P_n L(p_\epsilon) = \|P_n D^*(P_\epsilon)\|. \quad (6)$$

The one-step TMLE P_{ϵ_n} with $\epsilon_n = \arg \min_\epsilon P_n L(P_\epsilon)$, or ϵ_n chosen large enough so that the derivative is smaller than (e.g.) $1/n$, now solves $\frac{d}{d\epsilon} P_n L(P_\epsilon)|_{\epsilon=0} = 0$ (or $< 1/n$), and thus $\|P_n D^*(P_{\epsilon_n})\| = 0$ (or $< 1/n$). Note that $\|P_n D^*(P_{\epsilon_n})\| = 0$ implies that $P_n D_t^*(P_{\epsilon_n}) = 0$ for all $t \in \tau$ so that the one-step TMLE solves all desired estimating equations.

Table 2: Simulation Study II. Bias, variance, mean squared error (MSE) and relative efficiency (RE) of the one-step TMLE and iterative TMLE over 1000 Monte Carlo simulations, $n = 100$.

	Bias		Variance		MSE		RE
	One-Step	Iterative	One-Step	Iterative	One-Step	Iterative	
\bar{Q}_n unbounded, problematic runs omitted*							
Q_{c1}	0.00545	0.00544	0.00368	0.00367	0.00370	0.00370	1.00
Q_{c2}	0.00509	0.00512	0.00376	0.00376	0.00379	0.00379	1.00
Q_{c3}	0.00345	0.00346	0.00384	0.00384	0.00385	0.00385	1.00
Q_{c4}	0.00109	0.00110	0.00408	0.00408	0.00408	0.00408	1.00
Q_{c5}	-0.00047	-0.00031	0.00425	0.00425	0.00425	0.00425	1.00
Q_{c6}	-0.00276	-0.00279	0.00443	0.00444	0.00444	0.00445	1.00
Q_{c7}	-0.00468	-0.00472	0.00461	0.00462	0.00463	0.00464	1.00
Q_{c8}	-0.00821	-0.00873	0.00514	0.00550	0.00520	0.00557	0.93
Q_{c9}	-0.00980	-0.01013	0.00549	0.00561	0.00558	0.00570	0.98
Q_{c10}	-0.01306	-0.01311	0.00579	0.00580	0.00596	0.00597	1.00
\bar{Q}_n bounded at $(10^{-9}, 1 - 10^{-9})$							
Q_{c1}	-0.00308	-0.00687	0.00425	0.01649	0.00426	0.01652	0.26
Q_{c2}	-0.00341	-0.00672	0.00435	0.01509	0.00435	0.01512	0.29
Q_{c3}	-0.00480	-0.00979	0.00448	0.01363	0.00450	0.01371	0.33
Q_{c4}	-0.00622	-0.00909	0.00466	0.01508	0.00470	0.01515	0.31
Q_{c5}	-0.00784	-0.01420	0.00482	0.00872	0.00487	0.00891	0.55
Q_{c6}	-0.00958	-0.01613	0.00501	0.00708	0.00510	0.00734	0.69
Q_{c7}	-0.01117	-0.01519	0.00521	0.01002	0.00533	0.01024	0.52
Q_{c8}	-0.01393	-0.01979	0.00565	0.00777	0.00583	0.00816	0.72
Q_{c9}	-0.01528	-0.02086	0.00597	0.00778	0.00620	0.00821	0.76
Q_{c10}	-0.01746	-0.01745	0.00632	0.01241	0.00662	0.01270	0.52

*Number of omitted runs: 119, 110, 104, 94, 86, 73, 67, 57, 56, 42.

8.1 A universal canonical submodel that targets a multidimensional target parameter

Consider the following submodel: for $\epsilon \geq 0$, we define

$$\begin{aligned} p_\epsilon &= p \Pi_{[0, \epsilon]} \left(1 + \frac{\{P_n D^*(P_x)\}^\top D^*(P_x)}{\|D^*(P_x)\|} dx \right) \\ &= p \exp \left(\int_0^\epsilon \frac{\{P_n D^*(P_x)\}^\top D^*(P_x)}{\|D^*(P_x)\|} dx \right). \end{aligned} \quad (7)$$

Theorem 4 *We have $\{p_\epsilon : \epsilon \geq 0\}$ is a family of probability densities, its score at ϵ is a linear combination of $D_t^*(P_\epsilon)$ for $t \in \tau$, and is thus in the tangent space at $T(P_\epsilon)$, and*

$$\frac{d}{d\epsilon} P_n L(P_\epsilon) = \|P_n D^*(P_\epsilon)\|.$$

As a consequence, we have $\frac{d}{d\epsilon} P_n L(P_\epsilon) = 0$ implies $\|P_n D^(P_\epsilon)\| = 0$.*

As before, our practical construction below demonstrates that, under regularity conditions, we actually have that $\{p_\epsilon : \epsilon\} \subset \mathcal{M}$ is also a submodel.

The normalization by $\|D^*(P_x)\|$ is motivated by a practical analogue construction below and provides an important intuition behind this analytic construction. However, we can replace this by any other normalization for which the derivative of the log-likelihood at ϵ equals a norm of $P_n D^*(P_\epsilon)$. To illustrate this let's consider the case that $H = \mathbb{R}^d$. For example, we could consider the following submodel. Let $\Sigma_n(P_x) = P_n \{D^*(P_x) D^*(P_x)^\top\}$ be the empirical covariance matrix of $D^*(P_x)$, and let $\Sigma_n^{-1}(P_x)$ be its inverse. We could then define for $\epsilon > 0$,

$$p_\epsilon = p \exp \left(\int_0^\epsilon \{P_n D^*(P_x)\}^\top \Sigma_n^{-1} D^*(P_x) dx \right).$$

In this case, we have

$$\frac{d}{d\epsilon} P_n L(P_\epsilon) = P_n D^*(P_\epsilon)^\top \Sigma_n(P_\epsilon)^{-1} P_n D^*(P_\epsilon).$$

This seems to be an appropriately normalized norm, equal to the euclidean norm of the orthonormalized version of the original $D^*(P_\epsilon)$, so that the one-step TMLE will still satisfy that $\|P_n D^*(P_{\epsilon_n})\| = 0$.

It is not clear to us if these choices have a finite sample implication for the resulting one-step TMLE (asymptotics is the same), and if one choice would be better than another, but either way, the resulting one-step TMLE ends up with a P_{ϵ_n} satisfying $P_n D^*(P_{\epsilon_n}) = 0$ (or $o_P(1/\sqrt{n})$), the only key ingredient in the proof of the asymptotic efficiency of the TMLE.

8.2 The practical construction of a universal canonical one-dimensional submodel targeting a multidimensional target parameter

Let's define a local least favorable submodel $\{p_{\delta}^{\text{lfm}} : \delta\} \subset \mathcal{M}$ by the following local property: for all δ

$$\left. \frac{d}{d\delta} \log p_{\delta}^{\text{lfm}} \right|_{\delta=0}^{\top} \delta = D^*(P)^{\top} \delta.$$

For the case that $H = \mathbb{R}^d$, this corresponds with assuming that the score of the submodel at $\delta = 0$ equals the canonical gradient $D^*(P)$, while, for a general Hilbert space, it states that the derivative of $\log p_{\epsilon}$ in the direction δ (a function in H) equals $\langle D^*(P), \delta \rangle = \int D_t^*(P) \delta(t) d\Lambda(t)$.

Consider the log-likelihood criterion $P_n L(P_{\delta}^{\text{lfm}})$, and note that its derivative at $\delta = 0$ in the direction δ equals $\langle P_n D^*(P), \delta \rangle = \{P_n D^*(P)\}^{\top} \delta$. For a small number dx , we want to maximize the log-likelihood over all δ with $\|\delta\| \leq dx$, and locally, this corresponds with maximizing its linear gradient approximation:

$$\delta \rightarrow \{P_n D^*(P)\}^{\top} \delta.$$

By the Cauchy-Schwarz inequality, it follows that this is maximized over δ with $\|\delta\| \leq dx$ by

$$\delta_n^*(P, dx) = \frac{P_n D^*(P)}{\|P_n D^*(P)\|} dx \equiv \delta_n^*(P) dx,$$

where we defined $\delta_n^*(P) = P_n D^*(P) / \|P_n D^*(P)\|$. We can now define our update $P_{dx} = P_{\delta_n^*(P, dx)}^{\text{lfm}}$. This process can now be iterated by applying the above with P replaced by P_{dx} , resulting in an update P_{2dx} , and in general P_{Kdx} . So this updating process is defined by the differential equation:

$$P_{x+dx} = P_{x, \delta_n^*(P_x) dx}^{\text{lfm}},$$

where $P_{x, \delta}^{\text{lfm}}$ is the local least favorable multidimensional submodel above but now through P_x instead of P .

Assuming that the local least favorable model $h \rightarrow p_{x, h}^{\text{lfm}}$ is continuously twice differentiable with a score $D^*(P_x)$ at $h = 0$, we obtain a second order Taylor expansion

$$\begin{aligned} p_{x, \delta_n^*(P_x) dx}^{\text{lfm}} &= p_x + \left\{ \left. \frac{d}{dh} p_{x, h}^{\text{lfm}} \right|_{h=0} \right\}^{\top} \delta_n^*(P_x) dx + O((dx)^2) \\ &= p_x (1 + \{\delta_n^*(P_x)\}^{\top} D^*(P_x) dx) + O((dx)^2), \end{aligned}$$

so that, under mild regularity conditions, we obtain

$$p_{x+dx} = p_x (1 + \{\delta_n^*(P_x)\}^{\top} D^*(P_x) dx) + O((dx)^2).$$

This implies:

$$p_x = p \exp \left(\int_0^\epsilon \frac{\{P_n D^*(P_x)\}^\top}{\|P_n D^*(P_x)\|} D^*(P_x) dx \right).$$

So we obtained the exact same analytical representation (7) as above. Since the above practical construction starts out with $P \in \mathcal{M}$ and never leaves the model \mathcal{M} , this proves that, under mild regularity conditions, this analytic representation (7) is actually a submodel of \mathcal{M} after all, but, when using its practical implementation and approximation, one should use the actual local least favorable submodel in order to guarantee that one stays in the model. We can formalize this in a theorem analogue to Theorem 2, but instead such a theorem will be presented in Section 10 for the more general targeted minimum loss-based estimation methodology.

The above practical construction provides us with an intuition for the normalization by $\|P_n D^*(P_x)\|$.

8.3 Existence of MLE or approximate MLE ϵ_n .

Since

$$P_n \log p_\epsilon = \int_0^\epsilon \|P_n D^*(P_x)\| dx,$$

and its derivative thus equals $\|P_n D^*(P_\epsilon)\|$, we have that the log-likelihood is non-decreasing in ϵ .

If the local least favorable submodel in the practical construction of the one-dimensional universal canonical submodel $\{p_\epsilon : \epsilon \geq 0\}$ (7) only contains densities with supremum norm smaller than some $M < \infty$ (e.g., this is assumed by the model \mathcal{M}), then we will have that $\sup_{\epsilon \geq 0} \sup_{o \in \mathcal{O}} p_\epsilon(o) < M < \infty$. This implies that $P_n \log p_\epsilon$ is bounded from above by $\log M$. Let's first assume that $\lim_{\epsilon \rightarrow \infty} P_n \log p_\epsilon < \infty$. Thus, the log-likelihood is a strictly increasing function till it becomes flat, if ever. Suppose that $\limsup_{x \rightarrow \infty} \|P_n D^*(P_x)\| > \delta > 0$ for some $\delta > 0$. Then it follows that the log-likelihood converges to infinity when x converges to infinity, which contradicts the assumption that the log-likelihood is bounded from above by $\log M < \infty$. Thus, we know that $\limsup_{x \rightarrow \infty} \|P_n D^*(P_x)\| = 0$ so that we can find an ϵ_n so that for $\epsilon > \epsilon_n$ $\|P_n D^*(P_\epsilon)\| < 1/n$, as desired.

Suppose now that we are in a case in which the log-likelihood converges to infinity when $\epsilon \rightarrow \infty$, so that our bounded log likelihood assumption is violated. This might correspond with a case in which each p_ϵ is a continuous density, but p_ϵ starts approximating an empirical distribution when $\epsilon \rightarrow \infty$. Even in such a case, one would expect that we will have that $\|P_n D^*(P_\epsilon)\| \rightarrow 0$, just like an NPMLE of a continuous density of a survival time solves the efficient influence curve equation for its survival function.

The above practical construction of the submodel, as an iterative local maximization of the log-likelihood along its gradient, strongly suggests that even without the above boundedness assumption the derivative $\|P_n D^*(P_\epsilon)\|$ will converge to zero as

$\epsilon \rightarrow \infty$ so that the desired MLE or approximate MLE exists. Our initial practical implementations of this one-step TMLE of a multivariate target parameter demonstrate that it works well and that finding the desired maximum or approximate maximum is not an issue. We will demonstrate the implementation and practical demonstration of such a one-step TMLE for challenging causal inference problems in a future article.

8.4 A universal score-specific one-dimensional submodel targeting a multivariate score equation

In the above two subsections we could simply replace $D^*(P)$ by a user supplied $D(P)$, giving us a theoretical one-dimensional parametric model $\{P_\epsilon : \epsilon\}$ so that the derivative $\frac{d}{d\epsilon} P_n L(P_\epsilon)$ at ϵ equals $\|P_n D(P_\epsilon)\|$, so that a corresponding one-step TMLE will solve $P_n D(P_{\epsilon_n}) = 0$. Similarly, given a local parametric model whose score at $\epsilon = 0$ equals $D(P)$ will yield a corresponding practical construction of this universal submodel. One can also use such a universal score-specific submodel to construct one-step TMLE of a one-dimensional target parameter with extra properties by making it solve not only the efficient influence curve equation but also other equations of interest (such as the $P_n D_1^*(Q_n^*) = P_n D_2^*(Q_n^*) = 0$ in Section 6). In the current literature, solving multiple score equations typically required an iterative TMLE based on a local score-specific submodel, so that these estimation problems can be revisited with this new one-step TMLE (see our supplementary material).

9 Example: A one-step TMLE, based on universal canonical one-dimensional submodel, of an infinite dimensional target parameter

An open problem has been the construction of an efficient substitution estimator $\Psi(P_n^*)$ of a pathwise differentiable infinite dimensional target parameter $\Psi(P_0)$ such as a survival function. Current approaches would correspond with incompatible estimators such as using a TMLE for each $\Psi(P_0)(t)$ separately, resulting in a non-substitution estimator such as a non-monotone estimator of a survival function. In this section we demonstrate, through a causal inference example, that our universal canonical submodel allows us to solve this problem with the one-step TMLE defined in the previous section.

Let $O = (W, A, T) \sim P_0$, where W are baseline covariates, $A \in \{0, 1\}$ is a point-treatment, and T is a survival time. Consider a statistical model \mathcal{M} that only makes assumptions about the conditional distribution $g_0(a | W) = P_0(A = a | W)$ of A , given W . Let $W \rightarrow d(W) \in \{0, 1\}$ be a given dynamic treatment satisfying $g_0(d(W) | W) > 0$ a.e. Let $\Psi : \mathcal{M} \rightarrow H$ be defined by:

$$\Psi(P)(t) = E_P P(T > t | A = d(W), W), \quad t \geq 0.$$

Under a causal model and the randomization assumption this equals the counterfactual survival function $P(T_d > t)$ of the counterfactual survival time T_d under intervention d .

Let H be the Hilbert space of real valued functions on $\mathbb{R}_{\geq 0}$ endowed with inner product $h_1^\top h_2 = \langle h_1, h_2 \rangle = \int h_1(t)h_2(t)d\Lambda(t)$ for some user-supplied positive and finite measure Λ . The norm on this Hilbert space is thus given by $\|h\| = \sqrt{hh^\top} = \sqrt{\int h(t)^2 d\Lambda(t)}$. Let $\bar{Q}_t(A, W) = P(T > t \mid A, W)$, $Y(t) = I(T > t)$, Q_W the marginal probability distribution of W , and $Q = (\bar{Q}, Q_W)$. The efficient influence curve $D^*(P) = (D_t^*(P) : t \geq 0)$ is defined by:

$$\begin{aligned} D_t^*(P)(O) &= \frac{I(A = d(W))}{g(A \mid W)}(Y(t) - \bar{Q}_t(A, W)) + \{\bar{Q}_t(d(W), W) - \Psi(P)(t)\} \\ &\equiv D_{1,t}^*(g, \bar{Q}) + D_{2,t}^*(P), \end{aligned}$$

where $D_{1,t}^*(g, \bar{Q})$ is the first component of the efficient influence curve that is a score of the conditional distribution of T , given A, W . Notice that $\Psi(P) = \Psi_1(Q_W, \bar{Q}) = (Q_W \bar{Q}_t : t \geq 0)$. We will estimate $Q_{W,0}$ with the empirical distribution of W_1, \dots, W_n , so that a TMLE will only need to target the estimator of the conditional survival function \bar{Q}_0 of T , given A, W . Let $q(t \mid A, W)$ be the density of T , given A, W and let q_n be an initial estimator of this conditional density. For example, one might use machine learning to estimate the conditional hazard q_0/\bar{Q}_0 , which then implies a corresponding density estimator q_n . We are also given an estimator g_n of g_0 .

The universal canonical one-dimensional submodel (7) applied to $p = q_n$ is defined by the following recursive relation: for $\epsilon > 0$,

$$q_{n,\epsilon} = q_n \exp \left(\int_0^\epsilon \frac{\{P_n D_1^*(g_n, \bar{Q}_{n,x})\}^\top D_1^*(g_n, \bar{Q}_{n,x})}{\|D_1^*(g_n, \bar{Q}_{n,x})\|} dx \right).$$

To obtain some more insight in this expression, we note, for example, that the inner product is given by:

$$\{P_n D_1^*(g_n, \bar{Q}_{n,x})\}^\top D_1^*(g_n, \bar{Q}_{n,x})(o) = \int_t (P_n D_{1,t}^*(g_n, \bar{Q}_{n,x}) D_{1,t}^*(g_n, \bar{Q}_{n,x})(o) d\Lambda(t), \quad (8)$$

and similarly we have such an integral representation of the norm in the denominator. Our Theorem 4, or explicit verification, shows that for all $\epsilon \geq 0$, $q_{n,\epsilon}$ is a conditional density of T , given A, W , and

$$\frac{d}{d\epsilon} P_n \log q_{n,\epsilon} = \|P_n D_1^*(g_n, \bar{Q}_{n,\epsilon})\|.$$

Thus, if we move ϵ away from zero, the log-likelihood increases, and, one searches for the first ϵ_n so that this derivative is smaller than (e.g.) $1/n$. Let $q_n^* = q_{n,\epsilon_n}$, and let $\bar{Q}_{n,t}^*(A, W) = \int_t^\infty q_n^*(s \mid A, W) ds$ be its corresponding conditional survival

function, $t \geq 0$. Then our one-step TMLE of the d -specific survival function $\Psi(P_0)$ is given by $\psi_n^* = \Psi(Q_{W,n}, \bar{Q}_n^*) = Q_{W,n} \bar{Q}_n^*$:

$$\psi_n^*(t) = \frac{1}{n} \sum_{i=1}^n \bar{Q}_{n,t}^*(d(W_i), W_i).$$

Since q_n^* is an actual conditional density, it follows that ψ_n^* is a survival function. Suppose that the derivative of the log-likelihood at ϵ_n equals zero exactly (instead of being smaller than $1/n$). Then, we have $\|P_n D^*(g_n, Q_{W,n}, \bar{Q}_n^*)\| = 0$, so that for each $t \geq 0$, $P_n D_t^*(g_n, Q_{W,n}, \bar{Q}_n^*) = 0$, making $\psi_n^*(t)$ a standard TMLE of $\psi_0(t)$, so that its asymptotic linearity for a fixed t can be established accordingly. Let's now consider a proof of weak convergence of $\sqrt{n}(\psi_n^* - \psi_0)$ as a random function. Firstly, let's assume that an exact MLE is obtained so that $P_n D^*(g_n, Q_{W,n}, \bar{Q}_n^*) = 0$. Combined with $\Psi(Q_n^*) - \Psi(Q_0) = -P_0 D^*(g_n, Q_n^*) + R_2((Q_n^*, g_n), (Q_0, g_0))$, where $R_2() = (R_{2t}()) : t \in \tau$ for an explicitly defined $R_{2t}(P, P_0)$, we then obtain

$$\psi_n^* - \psi_0 = (P_n - P_0) D^*(g_n, Q_n^*) + R_2((Q_n^*, g_n), (Q_0, g_0)).$$

We now assume that $\{D_t^*(P) : P \in \mathcal{M}, t \in \tau\}$ is a P_0 -Donsker class, $\sup_{t \in \tau} P_0 \{D_t^*(g_n, Q_n^*) - D_t^*(g_0, Q_0)\}^2 \rightarrow 0$ in probability, and $\sup_t |R_{2t}((Q_n^*, g_n), (Q_0, g_0))| = o_P(n^{-1/2})$. Then, it follows that

$$\sqrt{n}(\psi_n^* - \psi_0) = \sqrt{n}(P_n - P_0) D^*(P_0) + o_P(n^{-1/2}) \Rightarrow_d G_0,$$

that is, $\sqrt{n}(\psi_n^* - \psi_0)$ converges weakly as a random element of the cadlag function space endowed with the supremum norm to a gaussian process G_0 with covariance structure implied by the covariance function $\rho(s, t) = P_0 D_s^*(P_0) D_t^*(P_0)$. In particular, if g_0 is known, then $R_{2t}((Q_n^*, g_0), (Q_0, g_0)) = 0$, so that the second order term condition $\sup_t |R_{2t}((Q_n^*, g_n), (Q_0, g_0))| = o_P(n^{-1/2})$ is automatically satisfied with $o_P(n^{-1/2})$ replaced by 0. This also allows the construction of a simultaneous confidence band for ψ_0 . Due to the double robustness of the efficient influence curve, under appropriate conditions, one can also obtain asymptotic linearity and weak convergence with an inefficient influence curve under misspecification of either g_n or \bar{Q}_n .

If we only have $\|P_n D^*(P_n^*)\| = o_P(n^{-1/2})$ (instead of 0), then the above proof still applies so that we now obtain:

$$\sqrt{n}(\psi_n^* - \psi_0) = (P_n - P_0) D^*(P_0) + r_n,$$

but where now $\|r_n\| = o_P(1/\sqrt{n})$, so that we obtain asymptotic efficiency and weak convergence in the Hilbert space $L^2(\Lambda)$, beyond the point-wise efficiency of $\psi_n^*(t)$. However, in practice, one can actually track the supremum norm $\|P_n D^*(P_{\epsilon_n})\|_\infty = \sup_t |P_n D_t^*(P_{\epsilon_n})|$, and if one observes that for the selected ϵ_n this supremum norm is smaller than $1/n$, then, we still obtain the asymptotic efficiency in supremum norm above.

Regarding the practical construction of $q_{n,\epsilon}$, we could use the following infinite dimensional local least favorable submodel through a conditional density q given by

$$q_{\delta}^{\text{lfm}} = q(1 + \delta^{\top} D_1^*(g, \bar{Q})),$$

and follow the practical construction described in the previous section for general local least favorable submodels. Notice that here $\delta^{\top} D_1^*(g, \bar{Q}) = \int \delta(t) D_{1,t}^*(g, \bar{Q}) d\Lambda(t)$. In order to guarantee that the supremum norm of the density q_{δ}^{lfm} for local δ with $\|\delta\| < dx$ remains below a universal constant $M < \infty$, one could present such models in the conditional hazard on a logistic scale that bounds the hazard between $[0, M]$. However, we doubt that this will be an issue in practice, and since it may be necessary for the continuous density $q_{n,\epsilon}$ to approximate an empirical distribution in some sense in order to solve $\|P_n D^*(P_{\epsilon})\| = 0$, we do not want to prevent this from happening.

10 Universal canonical one-dimensional submodel for targeted minimum loss-based estimation of a multi-dimensional target parameter

10.1 A universal canonical one-dimensional submodel

For the sake of presentation we will focus on the case that the target parameter is Euclidean valued, i.e. $H = \mathbb{R}^d$, but the presentation immediately generalizes to infinite dimensional target parameters, as in the previous section. Let's now generalize the construction of a universal canonical submodel to the more general targeted minimum loss based estimation methodology. We now assume that $\Psi(P) = \Psi_1(Q(P)) \in \mathbb{R}^d$ for some target parameter $Q : \mathcal{M} \rightarrow Q(\mathcal{M})$ defined on the model and real valued function $\Psi_1 : Q(\mathcal{M}) \rightarrow \mathbb{R}^d$. Let $L(Q)(O)$ be a loss-function for $Q(P)$ in the sense that $Q(P) = \arg \min_{Q \in Q(\mathcal{M})} PL(Q)$. Let $D^*(P) = D^*(Q(P), G(P))$ be the canonical gradient of Ψ at P , where $G : \mathcal{M} \rightarrow G(\mathcal{M})$ is some nuisance parameter. We consider the case that the linear span of the components of the efficient influence curve $D^*(P)$ is in the tangent space of Q , so that a least favorable submodel does not need to fluctuate G : otherwise, one just includes G in the definition of Q . Given, (Q, G) , let $\{Q_{\delta}^{\text{lfm}} : \delta\} \subset Q(\mathcal{M})$ be a local d -dimensional least favorable model w.r.t. loss function $L(Q)$ at $\delta = 0$ so that

$$\left. \frac{d}{d\delta} L(Q_{\delta}^{\text{lfm}}) \right|_{\delta=0} = D^*(Q, G).$$

The dependence of this submodel on G is suppressed in this notation.

Consider the empirical risk $P_n L(Q_{\delta}^{\text{lfm}})$, and note that its gradient at $\delta = 0$ equals $P_n D^*(Q, G)$. For a small number dx , we want to minimize the empirical risk over all

δ with $\|\delta\| \leq dx$, and locally, this corresponds with maximizing its linear gradient approximation:

$$\delta \rightarrow \{P_n D^*(Q, G)\}^\top \delta.$$

By the Cauchy-Schwarz inequality, it follows that this is maximized over δ with $\|\delta\| \leq dx$ by

$$\delta_n^*(Q, dx) = \frac{P_n D^*(Q, G)}{\|P_n D^*(Q, G)\|} dx \equiv \delta_n^*(Q) dx,$$

where we defined $\delta_n^*(Q) = P_n D^*(Q, G) / \|P_n D^*(Q, G)\|$. We can now define our update $Q_{dx} = Q_{\delta_n^*(Q, dx)}^{\text{lfm}}$. This process can now be iterated by applying the above with Q replaced by Q_{dx} , resulting in an update Q_{2dx} , and in general Q_{Kdx} . So this updating process is defined by the differential equation:

$$Q_{x+dx} = Q_{x, \delta_n^*(Q_x) dx}^{\text{lfm}},$$

where $Q_{x, \delta}^{\text{lfm}}$ is the local least favorable multidimensional submodel above but now through Q_x instead of Q .

Assume that for some $\dot{L}(Q)(O)$, we have

$$\left. \frac{d}{dh} L(Q_{x,h}^{\text{lfm}}) \right|_{h=0} = \dot{L}(Q_x) \left. \frac{d}{dh} Q_{x,h}^{\text{lfm}} \right|_{h=0}. \quad (9)$$

Then,

$$\left. \frac{d}{dh} Q_{x,h}^{\text{lfm}} \right|_{h=0} = \frac{D^*(Q_x, G)}{\dot{L}(Q_x)}.$$

Utilizing that the local least favorable model $h \rightarrow Q_{x,h}^{\text{lfm}}$ is continuously twice differentiable with a score $D^*(Q_x, G)$ at $h = 0$, we obtain a second order Taylor expansion

$$\begin{aligned} Q_{x, \delta_n^*(Q_x) dx}^{\text{lfm}} &= Q_x + \left. \frac{d}{dh} Q_{x,h}^{\text{lfm}} \right|_{h=0} \delta_n^*(Q_x) dx + O((dx)^2) \\ &= Q_x + \frac{D^*(Q_x, G)^\top}{\dot{L}(Q_x)} \delta_n^*(Q_x) dx + O((dx)^2). \end{aligned}$$

This implies the following recursive analytic definition of the universal canonical submodel through Q :

$$Q_\epsilon = Q + \int_0^\epsilon \frac{D^*(Q_x, G)^\top}{\dot{L}(Q_x)} \delta_n^*(Q_x) dx. \quad (10)$$

Let's now explicitly verify that this indeed satisfies the desired condition so that

the one-step TMLE solves $P_n D^*(Q_{\epsilon_n}, G) = 0$. Only assuming (9) it follows that

$$\begin{aligned}
\frac{d}{d\epsilon} P_n L(Q_\epsilon) &= P_n \frac{d}{d\epsilon} L(Q_\epsilon) \\
&= P_n \dot{L}(Q_\epsilon) \frac{d}{d\epsilon} Q_\epsilon \\
&= P_n \dot{L}(Q_\epsilon) \frac{D^*(Q_\epsilon, G)^\top}{\dot{L}(Q_\epsilon)} \delta_n^*(Q_\epsilon) \\
&= P_n D^*(Q_\epsilon, G)^\top \delta_n^*(Q_\epsilon) \\
&= \{P_n D^*(Q_\epsilon, G)\}^\top \frac{P_n D^*(Q_\epsilon, G)}{\|P_n D^*(Q_\epsilon, G)\|} \\
&= \frac{\sum_{j=1}^d \{P_n D_j^*(Q_\epsilon, G)\}^2}{\|P_n D^*(Q_\epsilon, G)\|} \\
&= \|P_n D^*(Q_\epsilon, G)\|.
\end{aligned}$$

In addition, under some regularity conditions, so that the following derivation in terms of the local least favorable submodel applies, it also follows that $Q_\epsilon \in Q(\mathcal{M})$.

This proves the following theorem.

Theorem 5 *Given any (Q, G) compatible with model \mathcal{M} , let $\{Q_\delta^{\text{lfm}} : \delta \in B_a(0)\} \subset Q(\mathcal{M})$ be a local least favorable model w.r.t. loss function $L(Q)$ at $\delta = 0$ so that*

$$\left. \frac{d}{d\delta} L(Q_\delta^{\text{lfm}}) \right|_{\delta=0} = D^*(Q, G).$$

Here $B_a(0) = \{x : \|x\| < a\}$ for some positive number a . Assume that for some $\dot{L}(Q)(O)$, we have

$$\left. \frac{d}{d\epsilon} L(Q_\epsilon^{\text{lfm}}) \right|_{\epsilon=0} = \dot{L}(Q) \left. \frac{d}{d\epsilon} Q_\epsilon^{\text{lfm}} \right|_{\epsilon=0}.$$

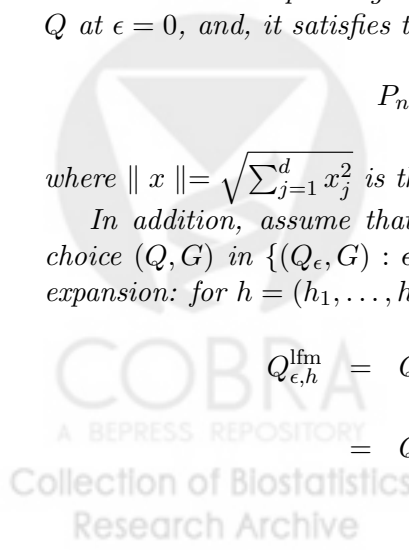
Consider the corresponding univariate model $\{Q_\epsilon : \epsilon\}$ defined by (10). It goes through Q at $\epsilon = 0$, and, it satisfies that for all ϵ

$$P_n \frac{d}{d\epsilon} L(Q_\epsilon) = \|P_n D^*(Q_\epsilon, G)\|, \quad (11)$$

where $\|x\| = \sqrt{\sum_{j=1}^d x_j^2}$ is the Euclidean norm.

In addition, assume that a in $B_a(0)$ can be chosen to be independent of the choice (Q, G) in $\{(Q_\epsilon, G) : \epsilon > 0\}$, and assume the following second order Taylor expansion: for $h = (h_1, \dots, h_d)$,

$$\begin{aligned}
Q_{\epsilon, h}^{\text{lfm}} &= Q_\epsilon + \left. \frac{d}{dh} Q_{\epsilon, h}^{\text{lfm}} \right|_{h=0} h + R_2(Q_\epsilon, G, \|h\|) \\
&= Q_\epsilon + \frac{D^*(Q_\epsilon, G)}{\dot{L}(Q_\epsilon)} h + R_2(Q_\epsilon, G, \|h\|),
\end{aligned}$$



where

$$\sup_{\epsilon} \sup_{o \in \mathcal{O}} |R_2(Q_{\epsilon}, G, \|h\|)(o)| = O(\|h\|^2).$$

We also assume that $\sup_{\epsilon} \sup_{o \in \mathcal{O}} \frac{|D^*(Q_{\epsilon}, G)|}{L(Q_{\epsilon})}(o) < \infty$.

Then, we also have $\{Q_{\epsilon} : \epsilon \geq 0\} \subset \mathcal{M}$.

11 Summary

Given a d -variate estimating function $(Q, O) \rightarrow D(Q, G)(O)$, a loss function $L(Q)$ for $Q : \mathcal{M} \rightarrow Q(\mathcal{M})$, a local d -dimensional submodel $\{Q_{\delta}^{sm} : \delta\} \subset Q(\mathcal{M})$ so that $\frac{d}{d\delta} L(Q_{\delta}^{sm})|_{\delta=0} = D(Q, G)$, we constructed a one-dimensional universal submodel $\{Q_{\epsilon} : \epsilon \geq 0\} \subset Q(\mathcal{M})$ through Q , at $\epsilon = 0$, that has the property that for all $\epsilon \geq 0$ $\frac{d}{d\epsilon} P_n L(Q_{\epsilon}) = \|P_n D(Q_{\epsilon}, G)\|$, where $\|\cdot\|$ is the Euclidean norm. Our analytic formula for this universal submodel does not depend on the local submodel, but the local submodel can still play a role for the practical construction. In the special case $d = 1$, we also constructed a universal one-dimensional submodel so that for all ϵ $\frac{d}{d\epsilon} L(Q_{\epsilon}) = D(Q_{\epsilon}, G)$, which then implies $\frac{d}{d\epsilon} P_n L(Q_{\epsilon}) = P_n D(Q_{\epsilon}, G)$. For each of these universal submodels, the one-step TMLE Q_{ϵ_n} with $\epsilon_n = \arg \min_{\epsilon} P_n L(Q_{\epsilon})$ solves each $P_n D_j(Q_{\epsilon_n}, G) = 0$, $j = 1, \dots, d$. We showed how this result immediately extends to an infinite dimensional estimating function $D = (D_t : t \in \tau)$, by replacing the Euclidean inner product by an Hilbert space inner product. If $D(\cdot)$ is the canonical gradient of a target parameter, we referred to this submodel as the universal canonical submodel, and, if $d = 1$, the universal least favorable submodel.

The constructions of these universal submodels correspond with iteratively defining $Q_{\epsilon+\delta\epsilon} = Q_{\epsilon, \delta(\epsilon)\delta\epsilon}^{sm}$ where $\delta(\epsilon) = P_n D(Q_{\epsilon}, G) / \|P_n D(Q_{\epsilon}, G)\|$ moves along the gradient of the empirical risk $P_n L(Q_{\epsilon})$ at ϵ . These practical constructions demonstrate that this algorithm succeeds in updating an initial Q into an update $Q_n^* = Q_{\epsilon_n}$ that solves the desired equation $P_n D(Q_{\epsilon_n}, G) = 0$ while *minimally* decreasing the empirical risk relative to its initial value $P_n L(Q)$. That is, with minimal additional data fitting it achieves the desired goal, while fully preserving the statistical properties of the initial estimator represented by Q .

The universal submodels have dramatic implications for the TMLE literature by allowing one to construct a one-step TMLE for any multivariate and even infinite dimensional pathwise differentiable target parameters, solving the desired estimating equation, so that this TMLE is asymptotically efficient and possibly has additional desired properties implied by solving the equation $P_n D(Q_{\epsilon_n}, G) = 0$. The one-step TMLE step only involves minimizing an empirical risk over a univariate fluctuation parameter ϵ . In the current literature, we defined various iterative TMLE based on multivariate local submodels that can now be replaced by a more stable one-step TMLE only relying on maximizing over a univariate ϵ . We demonstrated such new one-step TMLE for various examples in this article and supplementary material, but obviously this will impact many more problems than the ones presented here.

We demonstrated with a simulation study that the one-step TMLE is more robust and stable than the iterative TMLE in finite samples when the targeting step gets challenging.

The important advantages of the TMLE based on a local least favorable submodel relative to estimating equation methods and the one-step estimator have been emphasized in the literature. Since the estimating equation methodology is more limited than the one-step estimator by 1) relying on an estimating function representation of the efficient influence curve, 2) existence and 3) uniqueness of its solution, let's focus on contrasting the TMLE to the one-step estimator. One important advantage of the TMLE relative to the one-step estimator has been that it is a substitution estimator thereby making it in principle more robust by respecting the global constraints of the model \mathcal{M} . Beyond this, the fact that the TMLE updates an initial estimator through minimization of a loss-function specific empirical risk, it allows one to further refine the targeted update step such as carried out in C-TMLE. Another advantage is that it actually provides a corresponding data distribution $P_n^* \in \mathcal{M}$ compatible with the estimator of the target parameter, for example, allowing one to compare different TMLEs by the empirical risk of P_n^* . On the other hand, the one-step estimator takes only one step, and that can add important stability relative to a possibly iterative TMLE, making the comparison not so clear in the case that the TMLE is iterative. However, our new universal submodels presented in this article make the TMLE also a one step estimator, thereby dealing with this possible criticism of TMLE.

The benefit of being a substitution estimator is particularly appealing if one estimates an infinite dimensional target parameter such as a survival function with clear global structure. Due to our universal canonical one-dimensional submodel, we could provide one-step TMLE that completely respects this global structure of the infinite dimensional target parameter, something a one-step estimator (or estimating equation method) cannot achieve.

Future simulation studies will have to evaluate the practical benefits that come with the new one-step TMLEs based on universal least favorable or canonical submodels, relative to TMLEs based on the typical local least favorable submodel.

Acknowledgement

This grant is funded by NIH-grant 5R01AI074345-07. The authors also thank Marco Carone for stimulating discussions. We are particularly grateful to the two reviewers who have been very helpful.



References

- P.J. Bickel, C.A. Klassen, Y. Ritov, and J.A. Wellner. *Efficient and adaptive estimation of semiparametric models*. Johns Hopkins University Press, Baltimore, MD, 1993.
- M. Carone, I. Diaz, and M.J. van der Laan. Higher-order targeted minimum loss-based estimation. Technical Report 331, www.bepress.com/ucbbiostat/paper331, University of California, Berkeley, 2014.
- I. Diaz, M. Carone, and M.J. van der Laan. Second order inference for the mean of a variable missing at random. Technical Report 337, www.bepress.com/ucbbiostat/paper337, University of California, Berkeley, 2015.
- Iván Díaz and M.J. van der Laan. Targeted data adaptive estimation of the causal dose response curve. *Journal of Causal Inference*, 1(2), 2013.
- S. Gruber and M.J. van der Laan. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int J Biostat*, 6(1), 2010.
- S. Gruber and M.J. van der Laan. Targeted minimum loss based estimator that outperforms a given estimator. *The International Journal of Biostatistics*, 8(1): Article 11, doi: 10.1515/1557-4679.1332, 2012.
- S.D. Lendle, B. Fireman, and M.J. van der Laan. Balancing score adjusted targeted minimum loss-based estimation. *Journal of Causal Inference*, 3(2), 2015.
- E.C. Polley, S. Rose, and M.J. van der Laan. Super learning. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York Dordrecht Heidelberg London, 2012.
- J.M. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*, Methodological issues. Birkhäuser, 1992.
- D.B. Rubin and M.J. van der Laan. Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, Vol. 4, Iss. 1, Article 5, 2008.
- O.M. Stitelman and M.J. van der Laan. Collaborative targeted maximum likelihood for time to event data. Technical Report 260, Division of Biostatistics, University of California, Berkeley, 2010.
- M.J. van der Laan. Estimation based on case-control designs with known prevalence probability. *The International Journal of Biostatistics*, page <http://www.bepress.com/ijb/vol4/iss1/17/>, 2008.

- M.J. van der Laan. Statistical inference when using data adaptive estimators of nuisance parameters. Technical Report 302, Division of Biostatistics, University of California, Berkeley, submitted to IJB, 2012.
- M.J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical report, Division of Biostatistics, University of California, Berkeley, November 2003.
- M.J. van der Laan and S. Gruber. Collaborative double robust penalized targeted maximum likelihood estimation. *The International Journal of Biostatistics*, 2010.
- M.J. van der Laan and M.L. Petersen. Targeted learning. In *Ensemble Machine Learning*, chapter pages 117–156, ISBN 978-1-4419-9326-7. Springer, New York, 2012.
- M.J. van der Laan and J.M. Robins. *Unified methods for censored longitudinal data and causality*. Springer, New York, 2003.
- M.J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York, 2011.
- M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.
- M.J. van der Laan, S. Dudoit, and A.W. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics and Decisions*, 24(3):373–395, 2006.
- M.J. van der Laan, E. Polley, and A. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(25), 2007. ISSN 1.
- A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag New York, 1996.
- A.W. van der Vaart, S. Dudoit, and M.J. van der Laan. Oracle inequalities for multi-fold cross-validation. *Statistics and Decisions*, 24(3):351–371, 2006.
- W. Zheng and M.J. van der Laan. Cross-validated targeted minimum loss based estimation. In M.J. van der Laan and S. Rose, editors, *Targeted Learning: Causal Inference for Observational and Experimental Studies*. Springer, New York, 2011.
- W. Zheng, M.L. Petersen, and M.J. van der Laan. Estimating the effect of a community-based intervention with two communities. *Journal of Causal Inference*, 1(Issue 1):83–106, 2013.

Appendix

A Example for Section 4: Universal least favorable submodel for parametric models, and resulting one-step TMLE

This section represents the final subsection of Section 4.

Even though the standard MLE for a parametric model is asymptotically efficient for any pathwise differentiable target parameter, if the dimension of the finite dimensional parameter is high relative to sample size, then the MLE is often not well defined or overly variable so that regularization is needed, and in that case a TMLE is still needed. High dimensional linear regression is an example of such types of high dimensional parametric models, but also saturated models when O is discrete (but possibly with many possible values). This type of application of TMLE motivates us to consider the universal least favorable submodel and corresponding one-step TMLE for parametric models.

Let $O \sim P_{\theta_0} \in \mathcal{M} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ be modeled with a d -dimensional parametric model. Assume that the model is dominated by a single dominating measure μ . The density $dP_\theta/d\mu$ will be denoted with p_θ . Let $\Psi : \mathcal{M} \rightarrow \mathbb{R}$ be a real valued target parameter, which is pathwise differentiable with canonical gradient $D^*(P_\theta)$ at $P_\theta \in \mathcal{M}$. Let $S_j(P_\theta) = \frac{d}{d\theta_j} \log dP_\theta/d\mu$ be the score of θ_j , $j = 1, \dots, d$. The tangent space $T(P_\theta)$ at P_θ is the linear span of these d scores. Let $\alpha(P_\theta) = (\alpha_j(P_\theta) : j = 1, \dots, d)$ be the uniquely defined vector of scalars such that

$$D^*(P_\theta) = \sum_{j=1}^d \alpha_j(P_\theta) S_j(P_\theta).$$

Such a vector $\alpha(P_\theta)$ exists and is unique if the $d \times d$ information matrix $I(P_\theta) = P_\theta S_\theta S_\theta^\top$ is invertible, but even when the tangent space is of lower dimension than d , there exists a whole space of such vectors of scalars, and this just selects one of them in a unique manner.

A local least favorable model $\{P_{\theta,\epsilon}^{\text{lfm}} : \epsilon\}$ through P_θ at $\epsilon = 0$ is given by:

$$P_{\theta,\epsilon}^{\text{lfm}} = P_{\theta+\epsilon\alpha(P_\theta)} = P_{(\theta_j+\epsilon\alpha_j(P_\theta):j=1,\dots,d)}.$$

Let

$$\theta^{\text{lfm}}(\epsilon) = \theta + \epsilon\alpha(P_\theta)$$

be the corresponding least favorable path in the Θ space, so that we can denote

$P_{\theta,\epsilon}^{\text{lfm}} = P_{\theta^{\text{lfm}}(\epsilon)}$. Indeed,

$$\begin{aligned} \left. \frac{d}{d\epsilon} \log p_{\theta,\epsilon}^{\text{lfm}} \right|_{\epsilon=0} &= \left. \frac{1}{p_{\theta}} \frac{d}{d\epsilon} p_{\theta+\epsilon\alpha(P_{\theta})} \right|_{\epsilon=0} \\ &= \sum_{j=1}^d \frac{1}{p_{\theta}} \frac{d}{d\theta_j} p_{\theta} \frac{d}{d\epsilon} (\theta_j + \epsilon\alpha_j(P_{\theta})) \Big|_{\epsilon=0} \\ &= \sum_{j=1}^d \alpha_j(P_{\theta}) S_j(P_{\theta}) \\ &= D^*(P_{\theta}). \end{aligned}$$

Let the universal least favorable model through θ be defined by the following differential equation: for $\epsilon > 0, d\epsilon > 0$

$$\theta(\epsilon + d\epsilon) = \theta^{\text{lfm}}(\epsilon)(d\epsilon) = \theta(\epsilon) + d\epsilon\alpha(P_{\theta(\epsilon)}).$$

Similarly, we define $\theta(\epsilon - d\epsilon)$ for $\epsilon < 0$. The corresponding integral equation is given by: for $\epsilon > 0$ we have

$$\theta(\epsilon) = \theta + \int_0^{\epsilon} \alpha(P_{\theta(x)}) dx.$$

This differential or integral equation allows one to solve recursively for $\theta(\epsilon)$, given previous values $\theta(x)$ for $x < \epsilon$.

A corresponding universal least favorable submodel $\{P_{\theta,\epsilon} : \epsilon\}$ through P_{θ} is now defined by: for $\epsilon \geq 0$

$$\begin{aligned} P_{\theta,\epsilon} &= P_{\theta(\epsilon)} \\ &= P_{\theta + \int_0^{\epsilon} \alpha(P_{\theta(x)}) dx}. \end{aligned}$$

And similarly we can define $P_{\theta,\epsilon}$ for $\epsilon < 0$. By our results, we also know that we could define this universal least favorable submodel through P_{θ} by: for $\epsilon \geq 0$

$$P_{\theta,\epsilon} = P_{\theta} \exp \left(\int_0^{\epsilon} D^*(P_{\theta(x)}) dx \right),$$

but for the sake of practical approximation one should prefer the above formulation in terms of a local least favorable submodel.

So let's now discuss how one would implement the corresponding one-step TMLE. Let θ_n be an initial estimator. Suppose that $P_n \log p_{\theta_n^{\text{lfm}}(\epsilon)}$ is increasing at $\epsilon = 0$. Then, the TMLE is defined by defining ϵ_n as the smallest local maximum larger than 0 of $\epsilon \rightarrow P_n \log p_{\theta_n(\epsilon)}$, i.e., the log-likelihood along the universal least favorable submodel. The TMLE of θ_0 is now given by the one-step update $\theta_n^* = \theta_n(\epsilon_n)$, and the TMLE of $\Psi(P_{\theta_0})$ is given by $\Psi(P_{\theta_n^*})$.