Higher Order Targeted Maximum Likelihood Estimation in Causal Inference

Mark van der Laan
Division of Biostatistics, UC Berkeley

June 11-12, 2022, Washington D.C.
First International Workshop on Interactive Causal Learning

Acknowledgement: NIH R01AI074345

Reference: Mark van der Laan, Zeyi Wang, and Lars van der Laan (2021). Higher Order Targeted Maximum Likelihood Estimation. https://arxiv.org/abs/2101.06290

Outline

Outline

Statistical Estimation Problem

- We observe n i.i.d. copies O_i of $O \sim P_0$.
- We know $P_0 \in \mathcal{M}$ for a given statistical model $\mathcal{M}$.
- Let $\Psi : \mathcal{M} \to \mathbb{R}$ be the target parameter so that $\Psi(P_0)$ is the target estimand.
- Assume Ψ is pathwise differentiable at P with canonical gradient/efficient influence curve $D_P^{(1)}$: For each path $\{P_\delta : \delta\} \subset \mathcal{M}$ through P at $\delta = 0$ with score S we have

$$\frac{d}{d\delta}\Psi(P_\delta)\Big|_{\delta=0} = E_P D_P^{(1)}(O)S(O).$$

- We use notation $Pf \equiv \int f(o)dP(o)$ and $P_n f = 1/n \sum_i f(O_i)$.

TMLE

Initial Estimator: P_n^0 initial estimator of P_0.

Targeting of initial estimator: Construct so called universal least favorable parametric submodel $\{\tilde{P}^{(1)}(P_n^0, \epsilon) : \epsilon\} \subset \mathcal{M}$ through P_n^0 so that $\frac{d}{d\epsilon} L(P_{n,\epsilon}^0)$ spans the canonical gradient $D^{(1)}_{\tilde{P}^{(1)}(P_n^0,\epsilon)}$, where (e.g.) $L(P)(O) = -\log p(O)$ is log-likelihood loss. Let

$$\epsilon_n^{(1)} = \arg\min_{\epsilon} \sum_i L(\tilde{P}(P_n^0, \epsilon))(O_i)$$

be the MLE, and $P_n^{1,*} = P_{n,\epsilon_n}^0$.

TMLE of ψ_0: The TMLE of ψ_0 is plug-in estimator $\Psi(P_n^{1,*})$.

Solves efficient influence curve estimating equation:

$$P_n D^{(1)}_{P_n^{1,*}} \equiv \frac{1}{n} \sum_i D^{(1)}_{P_n^{1,*}}(O_i) = 0.$$

Analysis of first order TMLE

- Let $R^{(1)}(P, P_0) \equiv \Psi(P) - \Psi(P_0) - (P - P_0)D_P^{(1)}$ be the so called **exact second order remainder** for target parameter Ψ.
- Since $PD_P^{(1)} = 0$,

$$\Psi(P_n^{1,*}) - \Psi(P_0) = -P_0 D^{(1)}_{P_n^{1,*}} + R^{(1)}(P_n^{1,*}, P_0).$$

- Combined with score equation $P_n D^{(1)}_{P_n^{1,*}} = 0$, this yields the **key equation for a TMLE**:

$$\Psi(P_n^{1,*}) - \Psi(P_0) = (P_n - P_0) D^{(1)}_{P_n^{1,*}} + R^{(1)}(P_n^{1,*}, P_0).$$

- If $\underline{R^{(1)}(P_n^{1,*}, P_0) = o_P(n^{-1/2})}$ (and Donsker class condition), then

$$\Psi(P_n^{1,*}) - \Psi(P_0) = \frac{1}{n} \sum_{i=1}^{n} D^{(1)}_{P_0}(O_i) + o_P(n^{-1/2}).$$

- That is, ψ_n^* is asymptotically efficient plug-in estimator of ψ_0.

Example: Nonparametric estimation of treatment specific mean

- We observe n iid observations of $O = (W, A, Y) \sim P_0$, nonparametric model $\mathcal{M}$.
- W vector of baseline covariates, A binary treatment, Y binary outcome.
- $\Psi : \mathcal{M} \to \mathbb{R}$, $\Psi(P) = E_P E_P(Y \mid A = 1, W)$.
- Canonical gradient of pathwise derivative of Ψ at P:

$$D_P^{(1)} = \frac{A}{P(A|W)}(Y - E(Y \mid A, W)) + E(Y|A = 1, W) - \Psi(P).$$

- Exact second order remainder:

$$R^{(1)}(P, P_0) = P_0(\bar{Q} - \bar{Q}_0)(\bar{G} - \bar{G}_0)/\bar{G},$$

where $\bar{G}(W) \equiv P(A = 1 \mid W)$, $\bar{Q} \equiv E(Y|A = 1, W)$.

TMLE in ATE example

- The least favorable path through initial estimator $\bar{Q}_n^0$ is given by:

$$\text{Logit}\,\bar{Q}_{n,\epsilon}^0 = \text{Logit}\,\bar{Q}_n^0 + \epsilon A/\bar{G}_n,$$

- Let

$$\epsilon_n^1 = \arg\min_{\epsilon} P_n L(\bar{Q}_{n,\epsilon}^0)$$

be the MLE.

- Then, the TMLE of $\bar{Q}_0$ is given by the targeted initial estimator

$$\bar{Q}_n^* = \bar{Q}_{n,\epsilon_n^1}.$$

- The resulting plug-in TMLE of $\Psi(Q_0$ is thus:

$$\Psi(Q_n^*) = \frac{1}{n}\sum_{i=1}^{n} \bar{Q}_n^*(1, W_i),$$

utilizing an estimator $\bar{G}_n$ of propensity score.

Outline

The first order TMLE optimizes the exact total remainder of target parameter w.r.t. optimal empirical mean

- Define the exact total remainder
$$\bar{R}^{(0)}(P, P_0) = \Psi(P) - \Psi(P_0) - P_n D^{(1)}_{P_0}.$$
- Note
$$\bar{R}^{(0)}(P, P_0) = (P_n - P_0)\{D^{(1)}_P - D^{(1)}_{P_0}\} + R^{(1)}(P, P_0) - P_n D^{(1)}_P.$$
- Consider the oracle steepest descent algorithm optimizing $P \to \{\bar{R}^{(0)}(P, P_0)\}^2$ starting at $P = P^0_n$.
- The canonical gradient of this algorithm equals the canonical gradient $D^{(1)}_P$ of $\Psi(P)$ up till scalar $2(\Psi(P) - \Psi(P_0) - P_n D^{(1)}_{P_0})$ (telling which direction to move).
- Therefore, the TMLE algorithm (using universal LFM) $\arg\min_\epsilon P_n L(\tilde{P}^{(1)}(P, \epsilon)))$ is identical to this oracle steepest descent algorithm, except that it stops when the likelihood is maximized, at which point the sign of scalar is random.

Example demonstration of impact of TMLE-step

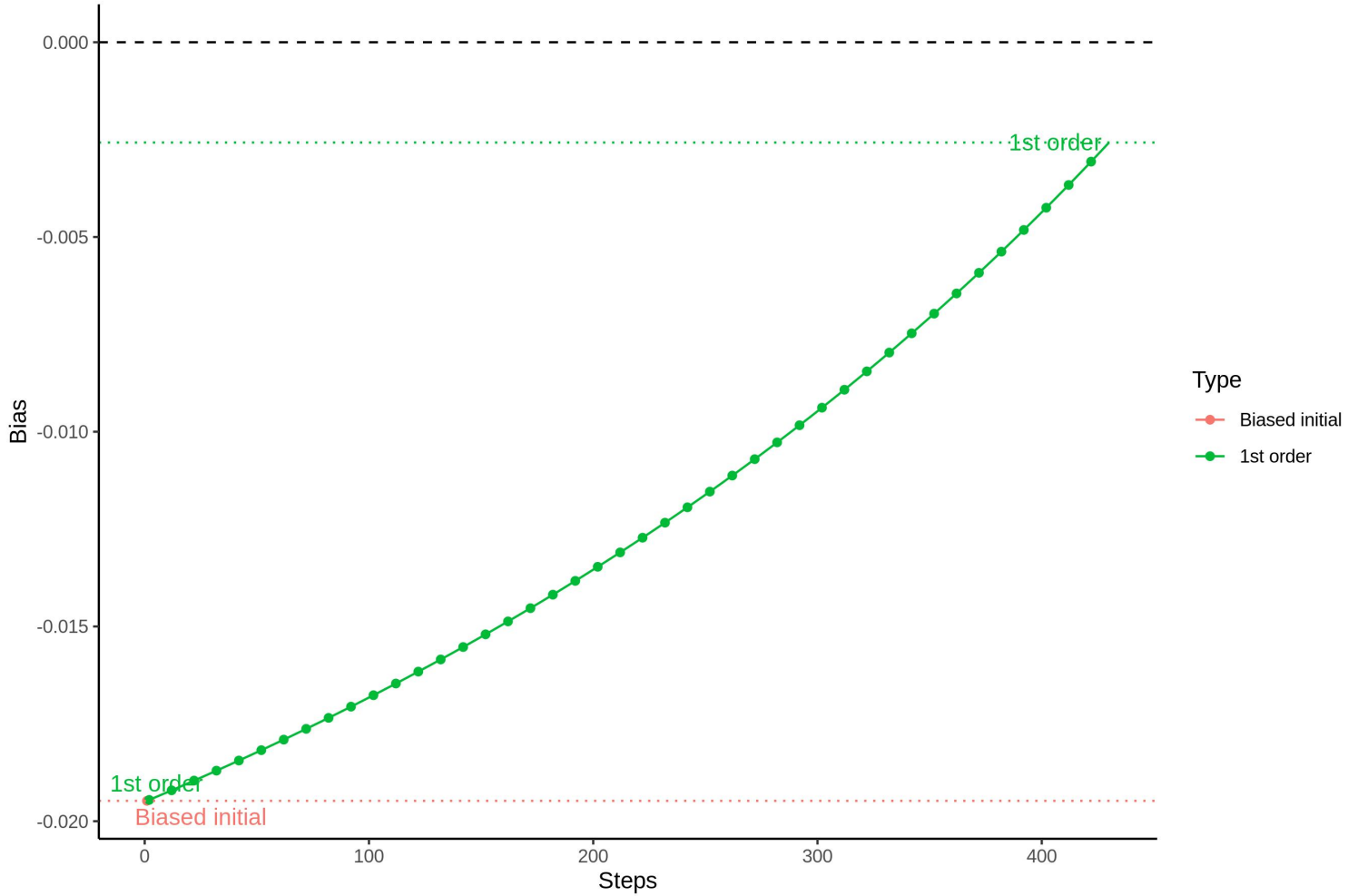
- Example: integrated squared density.
  - $\Psi : \mathcal{M} \to \mathbb{R}$, $\Psi(P) = \int p(x)^2 d\mu(o)$. $D_P^{(1)} = 2p - 2\Psi(P)$.
  - $R^{(1)}(P, P_0) = -\int (p - p_0)^2 d\mu$.
- Define TMLE updates of P_n^0 along the least favorable path:

$$\tilde{p}(P, \epsilon) = (1 + \epsilon D_P^{(1)})p.$$

$$\epsilon_n^{(1)} = \underset{\epsilon}{\text{argmin}} \; -P_n \log \tilde{p}(P_n^0, \epsilon).$$

Example demonstration of impact of TMLE-step

- Track TMLE steps $\epsilon \to \Psi(\tilde{P}(P_n^0, \epsilon)) - \Psi(P_0)$. $d\epsilon = 0.01$.



Outline

Representation of multivariate cadlag function

- A multivariate real valued cadlag function $f : [0,1]^d \to \mathbb{R}$ can be represented as

$$f(x) = \int_{[0,1]^d} \phi_u(x) df(u)$$

where

$$\phi_u(x) = I(x \geq u)$$

and $df(u)$ are the increments of the left-edge extended measure implied by f (thinking of f like a cumulative distribution function, lower dimensional on the left-edges).

- When components of knot-point u are zero, then the indicator basis function reduces to lower dimensional indicator basis function.
- The variation norm $\| f \|_v$ can be defined as $\| f \|_v = \int_{[0,1]^d} | df(u) |$.
- That is, f is an infinite linear combination of tensor product of indicator (zero-order spline) basis functions and the L_1-norm of its coefficients is the variation norm of f.

HAL-MLE

- Thus, for a given loss function $L(f, O)$ and functional parameter $f(P_0) = \arg\min_f P_0 L(f)$ with parameter space contained in $D[0,1]^d$, we can define a C-specific MLE

$$f_n^C = \arg \min_{f, \|f\|_v < C} P_n L(f).$$

- C can be selected with cross-validation.
- By using finite set of basis functions, $f_\beta = \sum_j \beta(j)\phi_{u_j}$, this becomes

$$\beta_n^C = \arg \min_{\beta, \|\beta\|_1 < C} P_n L(f_\beta),$$

and $f_n^C = f_{\beta_n^C}$.

- Therefore, we can fit any nuisance parameter function such as $\bar{Q}_0 = E(Y|A,W)$ and $\bar{G}_0 = E(A|W)$ needed for TMLE of ATE with Lasso-loss-based estimation using $n2^{d-1}$ basis functions, where the L_1-norm is selected with cross-validation.

Convergence rate of HAL-MLE, and efficiency of corresponding HAL-TMLE

- HAL-MLE f_n converges in loss-based dissimilarity/excess risk

$$d_0(f_n, f_0) = P_0 L(f_n) - P_0 L(f_0)$$

(like square of $L^2(P_0)$-norm) at a rate

$$n^{-1/3}(\log n)^{d/2})$$ Bibaut, vdL, 2019.

- As a consequence, a TMLE that uses as initial estimator an HAL-MLE $\tilde{P}_n$ (e.g., HAL-MLEs $\bar{Q}_n$ and $\bar{G}_n$ in ATE example) will be asymptotically efficient only assuming true nuisance functions are cadlag and of finite variation.

Example: *Asymptotic* efficiency of HAL-TMLE for treatment specific mean

Consider the HAL-TMLE of $EY_1 = EE(Y \mid A = 1, W)$ based on $(W, A, Y) \sim P_0$ in a nonparametric model.
It is asymptotically efficient if

- $\delta < P_0(A = 1 \mid W)$ for some $\delta > 0$;
- $W \to E_0(Y \mid A = 1, W)$ and $W \to P_0(A = 1 \mid W)$ are cadlag and have finite sectional variation norm.

Outline

Target initial estimator in TMLE towards its goal

- Let P represent the initial estimator so that TMLE $\tilde{P}_n^{(1)}(P) = \tilde{P}^{(1)}(P, \epsilon_n^{(1)}(P))$ implies substitution TMLE of $\Psi(P_0)$ given by

$$\Psi_n^{(1)}(P) \equiv \Psi(\tilde{P}_n^{(1)}(P)).$$

- $\Psi_n^{(1)}(P_0)$ can be viewed itself as a target parameter of P_0, we refer to as the first order fluctuated target parameter.
- It makes sense to use as initial estimator P in the first order TMLE the TMLE $\tilde{P}_n^{(2)}(P_n^0)$ of $\Psi_n^{(1)}(P_0)$, resulting in $\Psi(\tilde{P}_n^{(1)}\tilde{P}_n^{(2)}(P_n^0))$ of $\Psi(P_0)$.
- In this way, the initial estimator P is tailored/targeted towards its goal "estimate $\Psi_n^{(1)}(P_0)$ well".

Using regularized MLE-updates

- $P \to \Psi(\tilde{P}^{(1)}(P, \epsilon))$ is smooth in P, but $\epsilon_n^{(1)}(P) = \arg\min_\epsilon P_n L(\tilde{P}^{(1)}(P, \epsilon))$ is *not* pathwise differentiable due to dP_n/dP being infinite for typical $P \in \mathcal{M}$.
- Therefore we replace P_n in $\epsilon_n^{(1)}(P)$ by an HAL-MLE $\tilde{P}_n \in \mathcal{M}$ of P_0 and thus use

$$\tilde{\epsilon}_n^{(1)}(P) = \arg\min_\epsilon \tilde{P}_n L(\tilde{P}^{(1)}(P, \epsilon))$$

  instead.
- For components of P_n^0 that are fitted with empirical measure (i.e., NPMLE), we can set $\tilde{P}_n$-component to empirical measure as well: e.g., in ATE example, the marginal distribution of W under $\tilde{P}_n$ is kept at empirical.
- Now, $\Psi_n^{(1)} : \mathcal{M} \to \mathbb{R}$ is pathwise differentiable at P with canonical gradient $D_{n,P}^{(2)}$, and exact remainder $R_n^{(2)}(P, P_0) \equiv \Psi_n^{(1)}(P) - \Psi_n^{(1)}(P_0) + P_0 D_{n,P}^{(2)}$.

The second order TMLE

- A universal least favorable path $\tilde{P}_n^{(2)}(P, \epsilon)$ can be constructed, targeting $\Psi_n^{(1)}(P_0) = \Psi(\tilde{P}_n^{(1)}(P_0))$.
- This defines a TMLE $\tilde{P}_n^{(2)}(P) = \tilde{P}_n^{(2)}(P, \tilde{\epsilon}_n^{(2)}(P))$ with

$$\tilde{\epsilon}_n^{(2)}(P) = \arg\min_{\epsilon} \tilde{P}_n L(\tilde{P}_n^{(2)}(P, \epsilon)).$$

- It solves $\tilde{P}_n D^{(2)}_{n, \tilde{P}_n^{(2)}(P)} = 0$.
- The second order TMLE is defined as

$$P_n^{2,*} = \tilde{P}_n^{(1)} \tilde{P}_n^{(2)}(P_n^0)$$

with corresponding plug-in estimator

$$\Psi(\tilde{P}_n^{(1)} \tilde{P}_n^{(2)}(P_n^0)).$$

Outline

Defining the exact total remainder of first order TMLE in terms of initial P

- Define the exact total remainder for the first order TMLE $\Psi(\tilde{P}_n^{(1)}(P))$ with initial P as

$$\bar{R}^{(1)}(\tilde{P}_n^{(1)}(P), P_0) \equiv \Psi(\tilde{P}_n^{(1)}(P)) - \Psi(P_0) - (\tilde{P}_n - P_0)D^{(1)}_{\tilde{P}_n^{(1)} P_0}.$$

- Recall the first order fluctuated target parameter

$$\Psi_n^{(1)}(P) \equiv \Psi(\tilde{P}_n^{(1)}(P)).$$

- We have that exact total remainder as function of initial P equals $\Psi_n^{(1)}(P)$ up till a constant $C(P_0)$:

$$\bar{R}^{(1)}(\tilde{P}_n^{(1)}(P), P_0) = \Psi_n^{(1)}(P) - \Psi_n^{(1)}(P_0) + R^{(1)}(\tilde{P}_n^{(1)}(P_0), P_0).$$

Second order TMLE makes exact total remainder third order

By replacing the initial estimator P in the first order TMLE by this TMLE $\tilde{P}_n^{(2)}(P_n^0)$, which satisfies the key TMLE equation

$$\Psi_n^{(1)}(\tilde{P}_n^{(2)}P_n^0) - \Psi_n^{(1)}(P_0) = (\tilde{P}_n - P_0)D^{(2)}_{n,\tilde{P}_n^{(2)}(P_n^0)} + R_n^{(2)}(\tilde{P}_n^{(2)}(P_n^0), P_0),$$

the exact total remainder becomes

$$\begin{aligned}&\bar{R}^{(1)}(\tilde{P}_n^{(1)}\tilde{P}_n^{(2)}(P_n^0), P_0)\\ &= (\tilde{P}_n - P_0)D^{(2)}_{n,\tilde{P}_n^{(2)}(P_n^0)} + R_n^{(2)}(\tilde{P}_n^{(2)}(P_n^0), P_0) + R^{(1)}(\tilde{P}_n^{(1)}(P_0), P_0)(**),\end{aligned}$$

where $R_n^{(2)}$ will behave as third order difference.

Steepest descent algorithm minimizing exact total remainder in initial P is equivalent with computing second order TMLE-update

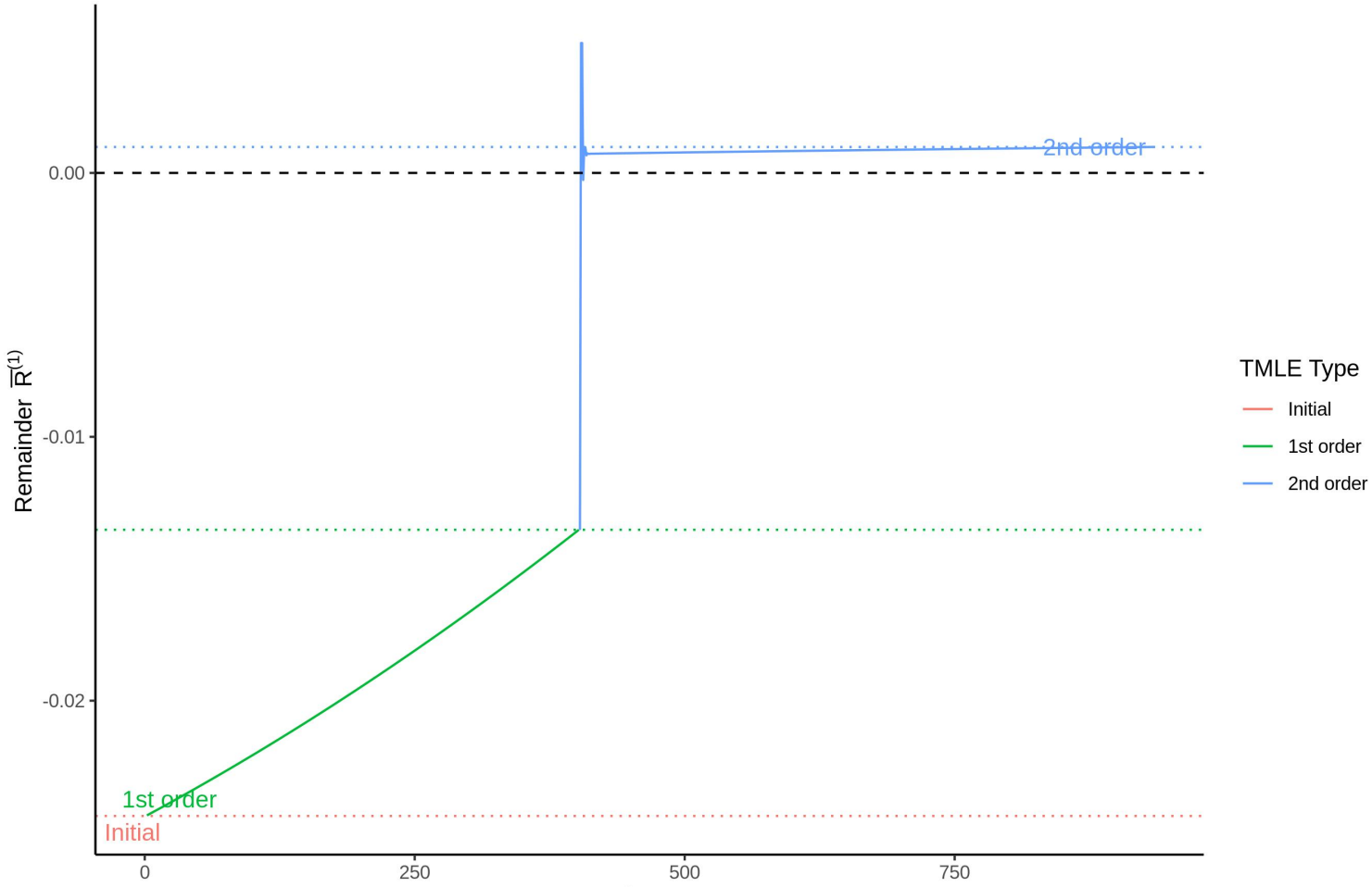
- To summarize:

$$\begin{aligned}\Psi(\tilde{P}_n^{(1)}(P)) - \Psi(P_0) &= (\tilde{P}_n - P_0)D^{(1)}_{\tilde{P}_n^{(1)}P_0} + \bar{R}^{(1)}(\tilde{P}_n^{(1)}(P), P_0) \\ &= (\tilde{P}_n - P_0)D^{(1)}_{\tilde{P}_n^{(1)}P_0} + \Psi_n^{(1)}(P) - \Psi_n^{(1)}(P_0) \\ &\quad + R^{(1)}(\tilde{P}_n^{(1)}(P_0), P_0).\end{aligned}$$

- Therefore to optimize the exact total remainder in initial P we want to set the initial P equal to a TMLE $\tilde{P}_n^{(2)}(P)$ targeting $\Psi_n^{(1)}(P)$.
- In fact, **running a (oracle) steepest descent algorithm on $\bar{R}^{(1)}(\tilde{P}_n^{(1)}P, P_0)$ is equivalent with computing the TMLE along the universal LFM $\tilde{P}_n^{(2)}(P, \epsilon)$ through P**, except the latter stops (early) when the likelihood cannot be increased anymore.

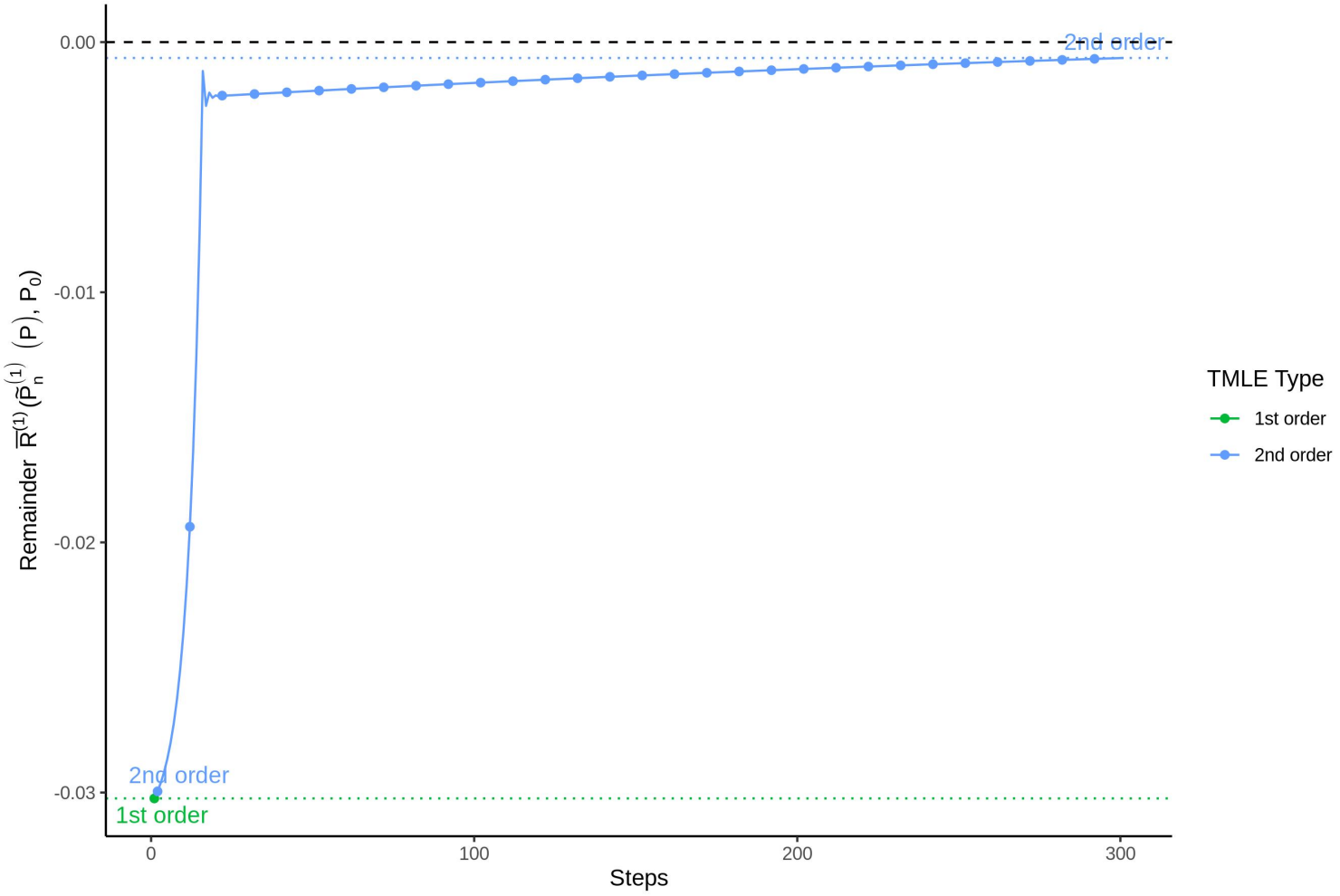
Demonstration of impact of second order TMLE-step

Below we plot $\epsilon \to \bar{R}^{(1)}(\tilde{P}_n^{(1)}\tilde{P}^{(2)}(P_n^0, \epsilon))$ along the universal LFM through P_n^0. As ϵ moves from 0 to its MLE $\tilde{\epsilon}_n^{(2)}(P_n^0)$, the moves from the exact remainder at standard first order TMLE to the exact remainder of second order TMLE.



Demonstration of impact of second order TMLE-step

- Even if p_n^0 equals constant $1/n$ and $\tilde{P}_n^{(1)}(P_n^0) = P_n^0$.



Outline

Exact expansion for the second order TMLE

- We have

$$\begin{aligned}
&\Psi(\tilde{P}_n^{(1)}\tilde{P}_n^{(2)}(P_n^0)) - \Psi(P_0) \\
&= (\tilde{P}_n - P_0)D^{(1)}_{\tilde{P}_n^{(1)}(P_0)} + R^{(1)}(\tilde{P}_n^{(1)}(P_0), P_0) \\
&+(\tilde{P}_n - P_0)D^{(2)}_{n,\tilde{P}_n^{(2)}(P_0)} + R_n^{(2)}(\tilde{P}_n^{(2)}(P_0), P_0) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad +\tilde{R}_n^{(3)},
\end{aligned}$$

  where

$$\tilde{R}_n^{(3)} = R_n^{(2)}(\tilde{P}_n^{(2)}(P_n^0), \tilde{P}_n) - R_n^{(2)}(\tilde{P}_n^{(2)}(P_0), \tilde{P}_n)$$

  is a difference of two third order differences.

- In fact, for these remainders at $\tilde{P}_n$ are iteratively targeted versions of $R^{(1)}()$:

$$R_n^{(2)}(\tilde{P}_n^{(2)}(P), \tilde{P}_n) = R^{(1)}(\tilde{P}_n^{(1)}\tilde{P}_n^{(2)}(P), \tilde{P}_n).$$

- Remainders $R^{(j)}(\tilde{P}_n^{(j)}(P_0), P_0)$, $j = 1, 2$, are practically negligible.

Undersmoothing the HAL-MLE to make MLE of ϵ behave as regular MLE

- By selecting an L_1-norm larger than the cross-validation selector in the HAL-MLE, the difference $\tilde{\epsilon}_n^{(1)}(P_0) - \epsilon_n^{(1)}(P_0)$ becomes small and negligible and thereby reduces size of **undersmoothing term**

$$(\tilde{P}_n - P_n)D^{(1)}_{\tilde{P}_n^{(1)}(P_0)} + (\tilde{P}_n - P_n)D^{(2)}_{n,\tilde{P}_n^{(2)}(P_0)} \approx 0.$$

- Then, the above exact expansion for the second order TMLE can be expressed in terms of

$$(P_n - P_0)D^{(1)}_{\tilde{P}_n^{(1)}(P_0)} + (P_n - P_0)D^{(2)}_{n,\tilde{P}_n^{(2)}(P_0)}.$$

Empirical higher order TMLE: Fitting MLEs with regular MLE

- In fact, we should just use the empirical MLEs $\epsilon_n^{(j)}(P)$, $j = 1, 2$, anyway, still using the HAL-MLE in the canonical gradient $D_{n,P}^{(2)}$.
- Then, the second order TMLE is nothing else than a TMLE targeting $\Psi(P_0)$ and the sequentially defined data adaptive fluctuation target parameters $\Psi_n^{(1)}(P_0)$.
- One can show this improves the exact expansion by having an improved undersmoothing term $(\tilde{P}_n - P_n)\{D^{(1)}_{\tilde{P}_n^{(1)}(P_0)} - D^{(1)}_{P_n^{2,*}}\}$, thereby less reliance on undersmoothing the HAL-MLE.

Undersmoothing term is third order difference $\sim n^{-1}$

- One can represent the undersmoothing term as $(\tilde{P}_n - P_n)f_n$ with f_n involving difference of P_n^0 and P_0.
- Consider initial estimator so that variation norm $\| f_n \|_v = O_P(r_1(n))$ converges to zero: e.g first order spline HAL with $r_1(n) = n^{-1/3}$.
- Define $\tilde{f}_n = f_n / \| f_n \|_v$ so that $(\tilde{P}_n - P_n)f_n = r_1(n)(\tilde{P}_n - P_n)\tilde{f}_n$.
- Since $\tilde{P}_n$ solves $P_n S_{j,\tilde{P}_n} = 0$ for a set of scores, we have

$$r_1(n)(\tilde{P}_n - P_n)\left\{\tilde{f}_n - \sum_j \alpha(j) S_{j,\tilde{P}_n}\right\} \equiv r_1(n)(\tilde{P}_n - P_n)e_n,$$

  where $\| e_n \|_{P_0} = O_P(n^{-1/3})$ (ignoring $\log n$-factors)
- Finally, write

$$r_1(n)(\tilde{P}_n - P_n)e_n = r_1(n)(\tilde{P}_n - P_0)e_n - r_1(n)(P_n - P_0)e_n = O_P(n^{-1}\log^m n)$$

- This rate does not reflect that $(\tilde{P}_n - P_n)$ can be controlled by undersmoothing.

Higher order inference

- Statistical inference can now be based on the second order expansion

$$\Psi(\tilde{P}_n^{2,*}) - \Psi(P_0) \approx (P_n - P_0) \left\{ D^{(1)}_{\tilde{P}_n^{(1)}(P_0)} + D^{(2)}_{n,\tilde{P}_n^{(2)}(P_0)} \right\} + \tilde{R}_n^{(3)}.$$

- The usual wald-type confidence interval can be constructed based on this sum estimated influence curve $\bar{D}_n = D_n^{(1)} + D_n^{(2)}$.
- One could also bootstrap this linear expansion $(P_n - P_0)\bar{D}_n$ to make the inference less reliant on normality.
- Inference now only ignores a third order remainder.

Outline

Generalization to k-th order TMLE

- For the naturally generalized k-th order TMLE $\tilde{P}_n^{(1)} \dots \tilde{P}_n^{(k)}(P_n^0)$ we have the analogue exact expansion:

$$\Psi(P_n^{k,*}) - \Psi(P_0) = \sum_{j=1}^{k} \left\{ (\tilde{P}_n - P_0) D^{(j)}_{n,\tilde{P}_n^{(j)}(P_0)} + R_n^{(j)}(\tilde{P}_n^{(j)}(P_0), P_0) \right\} + \tilde{R}_n^{(k+1)},$$

  where

$$\tilde{R}_n^{(k+1)} = R_n^{(k)}(\tilde{P}_n^{(k)}(P_n^0), \tilde{P}_n) - R_n^{(k)}(\tilde{P}_n^{(k)}(P_0), \tilde{P}_n)$$

  is a difference of two $k+1$-th order remainders.

- In fact,

$$R_n^{(k)}(\tilde{P}_n^{(k)}(P), \tilde{P}_n) = R^{(1)}(\tilde{P}_n^{(1)} \dots \tilde{P}_n^{(k)}(P), \tilde{P}_n).$$

- Again, using empirical TMLE updates instead of HAL-regularized ones only improves undersmoothing term.

Outline

Simulations for first and second order TMLE of the ATE

n	bias 1-st	bias 2-nd	se 1-st	se 2-nd	mse 1-st	mse 2-nd
400	**-0.720**	**0.078**	0.815	1.175	1.087	1.178
750	**-0.996**	**0.029**	0.800	1.102	1.278	1.102
1000	**-1.258**	**-0.062**	0.786	1.066	1.483	1.068
1200	**-1.345**	**0.022**	0.809	1.028	1.570	1.028
1600	**-1.549**	**-0.019**	0.818	1.055	1.752	1.055

Table: Simulation I: g_n^0 is $n^{-1/4}$-consistent, while Q_n^0 is inconsistent. The first order TMLE should have $n^{1/2}$-scaled bias that increases with n while the second order TMLE has a $n^{1/2}$-bias that should be constant in n. We observe that the second order TMLE has a negligible bias and thereby still provides valid inference.

Simulation II for ATE

n	bias 1-st	bias 2-nd	se 1-st	se 2-nd	mse 1-st	mse 2-nd
500	**-0.193**	**0.079**	0.858	1.062	0.879	1.065
1000	**-0.226**	**0.041**	0.942	1.126	0.968	1.126
1500	**-0.273**	**-0.022**	0.887	1.000	0.928	1.000
2500	**-0.244**	**0.027**	0.888	0.955	0.920	0.955
4000	**-0.256**	**0.077**	0.892	0.940	0.928	0.943

Table: Simulation II: g_n^0 and Q_n^0 are both $n^{-1/4}$-consistent. The first order TMLE should have $n^{1/2}$-scaled bias that does not converge to zero (but is constant in n), while the second order TMLE should have a $n^{-1/2}$-scaled bias that converges to zero at rate $n^{-1/4}$. We indeed observe that the second order TMLE has a negligible bias (bias/SE < 10), and thereby still provides valid inference.

Higher order TMLE when remainder is non-forgiving

- In double robust estimation problems, the remainder $R^{(1)}(P, P_0)$ has a cross-structure involving lot of cancelations.
- This often results in better finite sample behavior than one deserves based on the performance of the initial estimator P_n^0 of P_0.
- In many statistical estimation problems, including causal inference ones, the remainder $R^{(1)}(P, P_0)$ behaves as an integral of squared differences.
- For example, for a nonparametric model and target parameter $\Psi(P) = \int p^2 dx$, we have that $R^{(1)}(P, P_0) = -\int (p - p_0)^2(x) dx$.
- Such non-forgiving remainders will skew the finite sample distribution and make finite sample inference particularly hard.
- However, a **second order TMLE will replace this tough second order remainder not only by a third order remainder, but also one that involves lots of cancelation (no more squares)**.

Simulations for second order TMLE of integrated square density

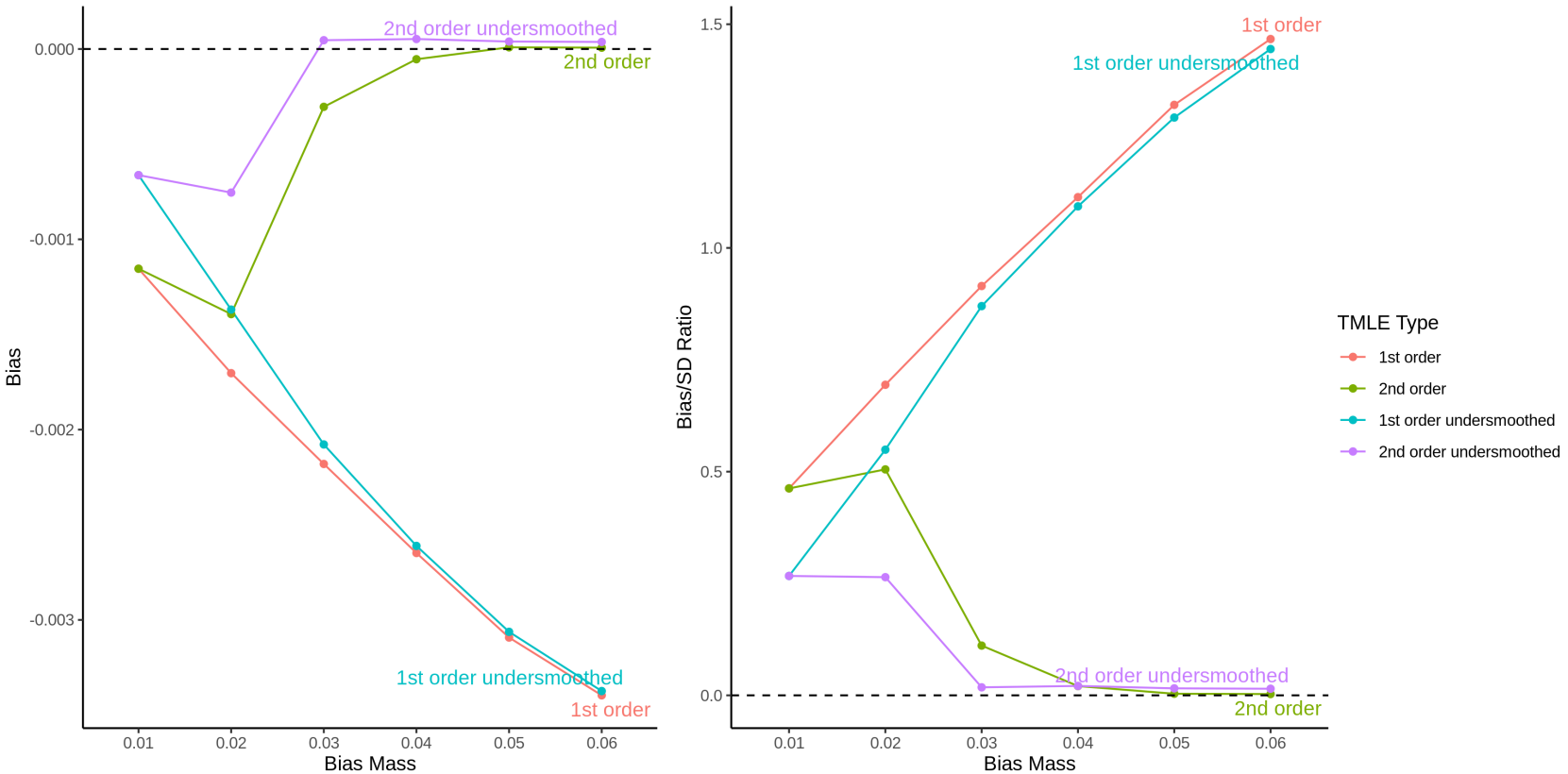


Figure: Bias (left) and Bias/SD ratio (right) performance. n=500. Second order TMLE provides additional total remainder control over first order TMLE following likelihood guidance in all scenarios, and remains consistent with increasingly biased initial P_n^0.

Simulations for second order TMLE of integrated square density

Table: Performance comparison. Initial P_n^0 is biased by adding point mass to empirical pmf supports and then scaling to sum 1. Second order TMLE controls first order TMLE exact total remainder. Undersmoothing controls $(\tilde{P}_n - P_n)D^{(1)}_{P_n^*}$ at the final TMLE update P_n^*.

Bias mass: 0.02	Bias	SD	MSE
1st order	**-1.70E-03**	2.45E-03	8.90E-06
2nd order	**-1.39E-03**	2.76E-03	9.50E-06
1st undersmoothed	-1.37E-03	2.49E-03	8.00E-06
2nd undersmoothed	**-7.54E-04**	2.86E-03	8.60E-06
Bias mass: 0.04			
1st order	**-2.65E-03**	2.38E-03	1.26E-05
2nd order	**-5.40E-05**	2.59E-03	6.70E-06
1st undersmoothed	-2.61E-03	2.39E-03	1.25E-05
2nd undersmoothed	**5.20E-05**	2.50E-03	6.20E-06

Outline

Concluding remarks

- k-th order TMLE allow linear expansion with a remainder that is a $k+1$-th order difference, beyond undersmooting term ($\sim 1/n$).
- Dramatic reductions in bias due to second order TMLE are observed in simulations.
- Not only weakens assumptions on initial estimator in TMLE for asymptotic efficiency, but has clear finite sample impact on bias and improves coverage of confidence intervals.
- Particularly appealing in estimation problems in which second order remainder behaves as square difference.
- Higher order TMLE also opens up a road for inference in problems in which the first order efficient influence curve disappears.
- Algebra and pure computational methods are developed for computing the higher order efficient influence curves and thereby higher order TMLE.