A Proof of Asymptotic Normality

Theorem 2. *Assume the following.*

1. *The mis-estimation of conditional outcomes can be bounded as follows*

$$\max_{a\in\{0,1\}} \mathbb{E}[(\hat{Q}_a(X)-Q(a,X))^2]^{\frac{1}{2}} = o(n^{-\frac{1}{4}}). \tag{4.9}$$

2. *The propensity score function $P(A=1|\cdot,\cdot)$ is Lipschitz continuous on $\mathbb{R}^2$, and $\exists\ \varepsilon>0$, $\mathrm{P}\left(\varepsilon \le g_\eta(X) \le 1-\varepsilon\right)=1$*

3. *The propensity score estimate converges at least as quickly as k nearest neighbor; i.e., $\mathbb{E}[\left(\hat{g}_\eta(X)-\mathrm{P}(A=1 \mid \hat{\eta}(X)\right)^2 \mid X]^{\frac{1}{2}} = O(n^{-\frac{1}{4}})$ Györfi et al. (2002);*

4. *There exist positive constants C_1, C_2, c, and $q>2$ such that*

$$\mathbb{E}[|Y|^q]^{\frac{1}{q}} \le C_2, \quad \sup_{\eta\in supp(\eta(X))} \mathbb{E}[(Y-Q(A,X)^2 \mid \eta(X)=\eta)] \le C_2,$$
$$\mathbb{E}[(Y-Q(A,X)^2)]^{\frac{1}{2}} \ge c, \quad \max_{a\in\{0,1\}} \mathbb{E}[\left|\hat{Q}_a(X)-Q(a,X)\right|]^{\frac{1}{q}} \le C_1.$$

Then, the estimator $\hat{\tau}^{TI}$ is consistent and

$$\sqrt{n}(\hat{\tau}^{TI}-\tau^{\mathrm{CDE}}) \xrightarrow{d} \mathbb{N}(0,\sigma^2) \tag{4.10}$$

where $\sigma^2 = E\left(\phi(X;Q,g_\eta,\tau^{\mathrm{CDE}})\right)^2$.

Proof. We first prove that misestimation of propensity score has rate $n^{-\frac{1}{4}}$. For simplicity, we use f_g, $\hat{f}_g$: $(u,v)\in\mathbb{R}^2 \to \mathbb{R}$ to denote conditional probability $\mathrm{P}(A=1|u,v) = f_g(u,v)$ and the estimated propensity function by running the nonparametric regression $\hat{\mathrm{P}}(A=1|u,v) = \hat{f}_g(u,v)$. Specifically, we have $f_g(Q(0,X),Q(1,X)) = g_\eta(X)$ and $\hat{f}_g(\hat{Q}_0(X),\hat{Q}_1(X)) = \hat{\mathrm{P}}(A=1|\hat{Q}_0(X),\hat{Q}_1(X)) = \hat{g}_\eta(X)$. Since $\mathbb{E}[\left(\hat{Q}_0(X)-Q(0,X)\right)^2]^{\frac{1}{2}}$, $\mathbb{E}[\left(\hat{Q}_1(X)-Q(1,X)\right)^2]^{\frac{1}{2}} = o(n^{-1/4})$ and f_g is Lipschitz continuous, we have

$$\begin{aligned}
&\mathbb{E}\left[\left|f_g(\hat{Q}_0(X),\hat{Q}_1(X))-f_g\left(Q(0,X),Q(1,X)\right)\right|^2\right]^{\frac{1}{2}} \\
\le& L\cdot\mathbb{E}\left[\left\|\left(\hat{Q}_0(X),\hat{Q}_1(X)\right)-\left(Q(0,X),Q(1,X)\right)\right\|_2^2\right]^{\frac{1}{2}} \\
=& L\cdot\left\{\mathbb{E}\left[\left(\hat{Q}_0(X)-Q(0,X)\right)^2\right]+\mathbb{E}\left[\left(\hat{Q}_1(X)-Q(1,X)\right)^2\right]\right\}^{\frac{1}{2}} \\
=& o(n^{-1/4})
\end{aligned} \tag{A.1}$$

Since the true propensity function f_g is Lipschitz continuous on $\mathbb{R}^2$, the mean squared error rate of the k nearest neighbor is $O(n^{-1/2})$ Györfi et al. (2002). In addition, since the propensity score function and its estimation are bounded under 1, we have the following equation

$$\mathbb{E}\left|\hat{f}_g(\hat{Q}_0(X),\hat{Q}_1(X))-f_g(\hat{Q}_0(X),\hat{Q}_1(X))\right|^2 = O(n^{-1/2}), \tag{A.2}$$

due to the dominated convergence theorem. By (A.1) and (A.2), we can bound the mean squared error of estimated propensity score in the following form:

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{g}_\eta(X)-g_\eta(X)\right)^2\right] \\
\leq&\mathbb{E}\left[\left(\hat{g}_\eta(X)-f_g(\hat{Q}_0(X),\hat{Q}_1(X))\right)^2\right]+\mathbb{E}\left[\left(f_g(\hat{Q}_0(X),\hat{Q}_1(X))-g_\eta(X)\right)^2\right] \\
=&\mathbb{E}\left|f_g(\hat{Q}_0(X),\hat{Q}_1(X))-f_g\left(Q(0,X),Q(1,X)\right)\right|^2+ \\
&\mathbb{E}\left|\hat{f}_g(\hat{Q}_0(X),\hat{Q}_1(X))-f_g(\hat{Q}_0(X),\hat{Q}_1(X))\right|^2 \\
=&O(n^{-1/2}),
\end{aligned} \tag{A.3}$$

that is $\mathbb{E}\left[\left(\hat{g}_\eta(X)-g_\eta(X)\right)^2\right]^{\frac{1}{2}}=O(n^{-\frac{1}{4}})$.

Before we apply the conclusion of Theorem 5.1 in (Chernozhukov et al., 2017b), we need to check all assumptions in Assumption 5.1 hold in Chernozhukov et al. (2017b). Let $C:=\max\left\{(2C_1^q+2^q)^{\frac{1}{q}},C_2\right\}$.

(a) $\mathbb{E}[Y-Q(A,X)\mid\eta(X),A]=0$, $\mathbb{E}[A-g_\eta(X)\mid\eta(X)]=0$ are easily checked by invoking definitions of Q and g_η.

(b) $\mathbb{E}[|Y|^q]^{\frac{1}{q}}\leq C$, $\mathbb{E}[(Y-Q(A,X))^2]^{\frac{1}{2}}\geq c$ and $\sup_{\eta\in\text{supp}(\eta(X))}\mathbb{E}[(Y-Q(A,X))^2\mid\eta(X)=\eta]\leq C$ are guaranteed by the fourth condition in the theorem.

(c) $\mathrm{P}\left(\varepsilon\leq g_\eta(X)\leq 1-\varepsilon\right)=1$ is the second condition in the theorem.

(d) Since propensity score function and its estimation are bounded under 1, we have

$$\begin{aligned}
&\left(\mathbb{E}[\left|\hat{Q}_1(X)-Q(1,X)\right|^q]+\mathbb{E}[\left|\hat{Q}_0(X)-Q(0,X)\right|^q]+\mathbb{E}[\left|\hat{g}_\eta(X)-g_\eta(X)\right|^q]\right)^{\frac{1}{q}} \\
&\leq(C_1^q+C_1^q+2^q)^{\frac{1}{q}} \\
&\leq C
\end{aligned}$$

(e) Based on (A.3) and condition 1 in the theorem, we have

$$\begin{aligned}
&\left(\mathbb{E}[\left(\hat{Q}_1(X)-Q(1,X)\right)^2]+\mathbb{E}[\left(\hat{Q}_0(X)-Q(0,X)\right)^2]+\mathbb{E}[\left(\hat{g}_\eta(X)-g_\eta(X)\right)^2]\right)^{\frac{1}{2}} \\
&\leq\left[o(n^{-\frac{1}{2}})+o(n^{-\frac{1}{2}})+O(n^{-\frac{1}{2}})\right]^{\frac{1}{2}} \\
&\leq O(n^{-\frac{1}{4}}), \\
&\mathbb{E}[\left(\hat{Q}_0(X)-Q(0,X)\right)^2]^{\frac{1}{2}}\cdot\mathbb{E}[\left(\hat{g}_\eta(X)-g_\eta(X)\right)^2]^{\frac{1}{2}}=o(n^{-\frac{1}{2}})
\end{aligned}$$

(f) Based on condition 3 in the theorem, we have

$$\sup_{x\in\text{supp}(X)}\mathbb{E}[\left(\hat{g}_\eta(X)-\mathrm{P}(A=1\mid\hat{\eta}(X))\right)^2\mid X=x]=O(n^{-\frac{1}{2}}).$$

We consider a smaller positive constant $\tilde{\varepsilon}$ instead of ε. Note that for $\tilde{\varepsilon} < \varepsilon$, we still have $\mathrm{P}(\tilde{\varepsilon} \le g_\eta(X) \le 1-\tilde{\varepsilon}) = 1$. Then,

$$
\begin{aligned}
&\mathrm{P}\left(\sup_{x\in\mathrm{supp}(X)}\left|\hat{g}_\eta(x)-\frac{1}{2}\right| > \frac{1}{2}-\tilde{\varepsilon}\right) = \mathrm{P}\left(\inf_{x\in\mathrm{supp}(X)}\hat{g}_\eta(x) < \tilde{\varepsilon}\right) + \mathrm{P}\left(\sup_{x\in\mathrm{supp}(X)}\hat{g}_\eta(x) > 1-\tilde{\varepsilon}\right) \\
\le& \mathrm{P}\left(\inf_{x\in\mathrm{supp}(X)}\mathrm{P}(A=1 \mid \hat{\eta}(X)=\hat{\eta}(x)) - \inf_{x\in\mathrm{supp}(X)}\hat{g}_\eta(x) > \varepsilon-\tilde{\varepsilon}\right) \\
&+ \mathrm{P}\left(\sup_{x\in\mathrm{supp}(X)}\hat{g}_\eta(x) - \sup_{x\in\mathrm{supp}(X)}\mathrm{P}(A=1 \mid \hat{\eta}(X)=\hat{\eta}(x)) > 1-\tilde{\varepsilon}-(1-\varepsilon)\right) \\
\le& \frac{\mathbb{E}\left[\left(\inf_{x\in\mathrm{supp}(X)}\hat{g}_\eta(x) - \inf_{x\in\mathrm{supp}(X)}\mathrm{P}(A=1 \mid \hat{\eta}(X)=\hat{\eta}(x))\right)^2\right]}{(\varepsilon-\tilde{\varepsilon})^2} + \\
&\frac{\mathbb{E}\left[\left(\sup_{x\in\mathrm{supp}(X)}\hat{g}_\eta(x) - \sup_{x\in\mathrm{supp}(X)}\mathrm{P}(A=1 \mid \hat{\eta}(X)=\hat{\eta}(x))\right)^2\right]}{(\varepsilon-\tilde{\varepsilon})^2} \\
\le& \frac{2\sup_{x\in\mathrm{supp}(X)}\mathbb{E}\left[\left(\hat{g}_\eta(X) - \mathrm{P}(A=1 \mid \hat{\eta}(X)=\hat{\eta}(x))\right)^2\right]}{(\varepsilon-\tilde{\varepsilon})^2} \\
=& O(n^{-\frac{1}{2}})
\end{aligned}
$$

Hence, $\mathrm{P}(\sup_{x\in\mathrm{supp}(X)}\left|\hat{g}_\eta(x)-\frac{1}{2}\right| \le \frac{1}{2}-\tilde{\varepsilon}) \ge 1-O(n^{-\frac{1}{2}})$.

With (a)-(f), we can invoke the conclusion in Theorem 5.1 in (Chernozhukov et al., 2017b), and get the asymptotic normality of the TI estimator. □

B Proof of Causal Identification

Theorem 1. *Assume the following:*
1. (Causal structure) The causal relationships among A, $\tilde{A}$, Z, Y, and X satisfy the causal DAG in Figure 2;
2. (Overlap) $0 < \mathrm{P}(A=1|X_{A\wedge Z}, X_Z) < 1$*;*
3. (Intention equals perception) $A = \tilde{A}$ *almost surely with respect to all interventional distributions. Then, the CDE is identified from observational data as*

$$
\mathrm{CDE} = \tau^{\mathrm{CDE}} := \mathbb{E}_{X|\tilde{A}=1}\left[\mathbb{E}[Y \mid \eta(X), \tilde{A}=1] - \mathbb{E}[Y \mid \eta(X), \tilde{A}=0]\right], \tag{3.4}
$$

where $\eta(X) := (Q(0,X), Q(1,X))$.

Proof. We first prove that this two-dimensional confounding part $\eta(X)$ satisfies positivity. Since $(Q(0,X),\ Q(1,X)) = (\mathbb{E}[Y \mid\ A=1, X_{A\wedge Z}, X_Z],\ \mathbb{E}[Y \mid\ A=0, X_{A\wedge Z}, X_Z])$ is a function of $(X_{A\wedge Z},\ X_Z)$, the following equations hold:

$$
\begin{aligned}
\mathrm{P}(A=1 \mid Q(0,X), Q(1,X)) =& \mathbb{E}(A \mid Q(0,X), Q(1,X)) \\
=& \mathbb{E}\left[E(A \mid X_{A\wedge Z}, X_Z) \mid Q(0,X), Q(1,X)\right] \\
=& \mathbb{E}\left[\mathrm{P}(A=1 \mid X_{A\wedge Z}, X_Z) \mid Q(0,X), Q(1,X)\right].
\end{aligned} \tag{B.1}
$$

As $0 < \mathrm{P}(A=1 \mid X_{A\wedge Z}, X_Z) < 1$, we have $0 < \mathrm{P}(A=1 \mid Q(0,X), Q(1,X)) < 1$. Furthermore, we have $0 < \mathrm{P}(\tilde{A}=1 \mid Q(0,X), Q(1,X)) < 1$ due to almost everywhere equivalence of A and $\tilde{A}$.

Since $A = \tilde{A}$, we can rewrite (3.1) by replacing A with $\tilde{A}$ in the following form:

$$\begin{aligned}
\text{CDE} =& \mathbb{E}_{X_{A\wedge Z}, X_Z \mid \tilde{A}=1}\big[\ \mathbb{E}(Y \mid \text{do}(\tilde{A}=1), X_{A\wedge Z}, X_Z) - \mathbb{E}(Y \mid \text{do}(\tilde{A}=0), X_{A\wedge Z}, X_Z)\ \big] \\
=& \mathbb{E}_{X_{A\wedge Z}, X_Z \mid \tilde{A}=1}\big[\ \mathbb{E}(Y \mid \tilde{A}=1, X_{A\wedge Z}, X_Z) - \mathbb{E}(Y \mid \tilde{A}=0, X_{A\wedge Z}, X_Z)\ \big] \\
=& \mathbb{E}_{X_{A\wedge Z}, X_Z \mid \tilde{A}=1}\big[\ \mathbb{E}(Y \mid \tilde{A}=1, X) - \mathbb{E}(Y \mid \tilde{A}=0, X)\ \big] \\
=& \mathbb{E}_{X_{A\wedge Z}, X_Z \mid \tilde{A}=1}\big[\mathbb{E}(Y \mid \tilde{A}=1, Q(0,X), Q(1,X))\big] - \mathbb{E}\big[\mathbb{E}(Y \mid \tilde{A}=0, Q(0,X), Q(1,X))\big] \\
=& \mathbb{E}_{X_{A\wedge Z}, X_Z \mid \tilde{A}=1}\big[\mathbb{E}(Y \mid \tilde{A}=1, \eta(X))\big] - \mathbb{E}\big[\mathbb{E}(Y \mid \tilde{A}=0, \eta(X))\big] \\
=& \mathbb{E}_{X \mid \tilde{A}=1}\big[\mathbb{E}(Y \mid \tilde{A}=1, \eta(X))\big] - \mathbb{E}\big[\mathbb{E}(Y \mid \tilde{A}=0, \eta(X))\big].
\end{aligned} \tag{B.2}$$

The equivalence of the first and the second line is because $X_{A\wedge Z}$, X_Z block all backdoor paths between $\tilde{A}$ and Y (See Figure 2) and $0 < \mathrm{P}(\tilde{A}=1 \mid Q(0,X), Q(1,X)) < 1$. Thus, the "*do-operation*" in the first line can be safely removed. Equivalence of the second line and the third line is due to $Q(\tilde{A}, X) = \mathbb{E}\left(Y \mid \tilde{A}, X_{A\wedge Z}, X_Z\right)$, which is subject to the causal model in Figure 2. The last equation is based on the fact that $\eta(X)$ is a function of only $X_{A\wedge Z}$ and X_Z. (It can be easily checked by using the definition of the expectation.)

(B.2) shows that $(Q(0,X), Q(1,X))$ is a two-dimensional confounding variable such that CDE is identifiable when we adjust for it as the confounding part.

□

Note that if f and h are two invertible functions on $\mathbb{R}$, $(f(Q(0,X)), h(Q(1,X)))$ also suffices the identification for CDE. Since the sigma algebra should be the same for $(Q(0,X), Q(1,X))$ and $f(Q(0,X)), h(Q(1,X))$, i.e.,

$$\sigma(Q(0,X), Q(1,X)) = \sigma(f(Q(0,X)), h(Q(1,X))).$$

Hence, we have

$$\begin{aligned}
&\mathrm{P}(A=1 \mid Q(0,X), Q(1,X)) = \mathrm{P}(A=1 \mid f(Q(0,X)), h(Q(1,X))), \\
&\mathbb{E}(Y \mid Q(0,X), Q(1,X)) = \mathbb{E}(Y \mid f(Q(0,X)), h(Q(1,X))).
\end{aligned} \tag{B.3}$$

C Additional Experiments

We conduct additional experiments to show how the estimation of causal effect changes 1) over different nonparametric models for the propensity score estimation, and 2) when using different double machine learning estimators on causal estimation. Specifically, for the first study, we apply different nonparametric models and the logistic regression to the estimated confounding part $\hat{\eta}(X) = \left(\hat{Q}_0(X), \hat{Q}_1(X)\right)$ to obtain propensity scores. We use ATT AIPTW in all above cases for causal effect estimation. For the second study, we fix the first two stages of the TI estimator, i.e. we apply Q-Net for the conditional outcomes and compute propensity scores with the Gaussian process regression where the kernel function is the summation of dot product and white noise. Estimated conditional outcomes and propensity scores are plugged into different double machine learning estimators. We make the following conclusions with results of above experiments.

The choice of nonparametric models is significant. Table 3 summarizes results with applying different regression models for the propensity estimation. We can see that suitable nonparametric models will strongly increase the coverage proportion over true causal estimand. Therefore, we conclude that the accuracy in causal estimation is highly dependent on the choice of nonparametric models. In practice, when there is some prior information about the propensity score function, we should apply the most suitable nonparametric model to increase the reliability of our causal estimation.

The ATT AIPTW is consistently the best double machine learning estimator. Table 4 shows results by applying different double machine learning estimators. We apply both estimators for the average treatment effect (ATE) and the controlled direct effect (CDE). The bias of "unadjusted" estimator $\hat{\tau}^{\text{naive}}$ is also included in Table 4 (a). For bias, ATT AIPTW

$\hat{\tau}^{\text{TI}}$ has comparable results with other double machine learning estimators in most cases. For coverage proportion of confidence intervals, though it has lower rates in some cases, $\hat{\tau}^{\text{TI}}$ has consistently the best performance. Especially in high confounding situations, the advantage of $\hat{\tau}^{\text{TI}}$ is obvious.

Estimator For each dataset, we compute estimators as follows. n_1 and n_0 stands for the number of individuals in the treated and controlled group. $n = n_1 + n_0$ is the total number of individuals.

- "Unadjusted" baseline estimator: $\hat{\tau}^{\text{naive}} = \frac{1}{n_1}\sum_{i:A_i=1} Y_i - \frac{1}{n_0}\sum_{i:A_i=0} Y_i$
- "Outcome-only" estimator: $\hat{\tau}^{\text{Q}} = \frac{1}{n_1}\sum_{i:A_i=1} \hat{Q}_{1,i} - \hat{Q}_{0,i}$
- ATT AIPTW: $\hat{\tau}^{\text{TI}} = \frac{1}{n_1}\sum_{i:A_i=1} A_i(Y_i - \hat{Q}_{0,i}) - (1-A_i)(Y_i - \hat{Q}_{0,i})\frac{\hat{g}_i}{1-\hat{g}_i}$

Table 3: The choice of nonparametric models for the TI-estimator is significant. Tables show average bias and 95% confidence intervals' coverage of $\hat{\tau}^{\text{TI}}$ with applying different nonparametric models in the second stage. The Gaussian process regression with the dot product+ white noise kernel has the best performance (lowest bias and highest coverage proportion). The treatment level is equal to true CDE, which takes 1.0 (with causal effect) and 0.0 (without causal effect). Low and high noise level corresponds to $\gamma = 1.0$ and 4.0. Low and high confounding level corresponds to $\beta_c = 50.0$ and 100.0.

(a) Average bias

Noise:	Low				High			
Treatment (oracle causal effect):	1.0		0.0		1.0		0.0	
Confounding:	Low	High	Low	High	Low	High	Low	High
GPR (Dot Product+White Noise)	0.069	0.059	0.113	0.074	0.088	0.049	0.002	0.089
GPR (RBF)	0.150	0.348	0.156	0.329	0.363	0.452	0.344	0.424
KNN	0.147	0.334	0.144	0.313	0.316	0.372	0.304	0.356
AdaBoost	0.074	0.349	0.061	0.323	0.526	0.497	0.479	0.464
Logistic	0.070	0.057	0.114	0.073	0.086	0.047	-0.001	0.087

(b) Coverage proportions of 95% confidence intervals

Noise:	Low				High			
Treatment (oracle causal effect):	1.0		0.0		1.0		0.0	
Confounding:	Low	High	Low	High	Low	High	Low	High
GPR (Dot Product+White Noise)	57%	84%	57%	79%	87%	80%	77%	81%
GPR (RBF)	31%	0%	41%	0%	7%	7%	17%	19%
KNN	18%	0%	39%	0%	11%	8%	11%	8%
AdaBoost	25%	0%	35%	0%	0%	0%	0%	0%
Logistic	58%	84%	57%	79%	87%	80%	78%	81%

D Discussion of Low Coverage

In this section, we discuss why the confidence intervals we get (See Table 1) have lower coverage than the nominated level 95%. We conduct diagnostics and find that the inaccuracy of Q's estimations is responsible for the low coverage. We compute biases, variances, and coverages of τ^{TI}'s with different mean squared errors $\hat{\mathbb{E}}[(Q-\hat{Q})^2]$ by using different numbers of datasets. According to Figure 4–Figure 5, as the mean squared error of Q increases, the bias of τ^{TI} grows and the coverage of τ^{TI} drops. Specifically, the highest coverage of each setting is almost 95% (use 50 datasets with most accurate conditional outcome estimations). In practice, one direct way to improve the TI estimator's accuracy is to apply better NLP models so that more accurate conditional outcome estimations can be obtained.

Table 4: The ATT AIPTW is consistently the best double machine learning estimator for this causal problem. Tables show average bias and 95% confidence intervals' coverage of different causal estimations. ATT AIPTW $\hat{\tau}^{TI}$ shows consistently the lowest bias and highest coverage rate. For propensity score estimation, the Gaussian process regression with the dot product+ white noise kernel is applied for all estimators. The treatment level is equal to true CDE/true ATE, which takes 1.0 (with causal effect) and 0.0 (without causal effect). Low and high noise level corresponds to $\gamma = 1.0$ and 4.0. Low and high confounding level corresponds to $\beta_c = 50.0$ and 100.0.

(a) Average bias

Noise:	Low				High			
Treatment (oracle CDE):	1.0		0.0		1.0		0.0	
Confounding:	Low	High	Low	High	Low	High	Low	High
unadjusted $\hat{\tau}^{naive}$	1.071	2.143	1.071	2.1453	1.068	2.140	1.069	2.140
ATE AIPTW	0.094	0.178	0.128	0.195	0.122	0.106	0.061	0.140
ATE BMM	0.094	0.176	0.128	0.193	0.122	0.106	0.061	0.140
ATE IPTW	-0.574	-1.492	-1.839	-1.807	-0.082	-0.592	-0.393	-0.649
ATT AIPTW: $\hat{\tau}^{TI}$	0.069	0.059	0.114	0.074	0.088	0.049	0.002	0.089
ATT BMM	0.075	0.147	-0.031	0.062	0.621	0.454	0.464	0.337
ATT TMLE:	0.084	0.194	0.085	0.196	0.186	0.136	0.174	0.163

(b) Coverage Proportions of 95% confidence intervals

Noise:	Low				High			
Treatment (oracle CDE):	1.0		0.0		1.0		0.0	
Confounding:	Low	High	Low	High	Low	High	Low	High
ATE AIPTW	37%	36%	69%	33%	75%	79%	79%	71%
ATE BMM	39%	35%	70%	36%	75%	79%	79%	71%
ATE IPTW	11%	1%	0%	1%	90%	39%	44%	37%
ATT AIPTW: $\hat{\tau}^{TI}$	57%	84%	57%	79%	87%	80%	77%	81%
ATT BMM	26%	4%	49%	41%	1%	3%	1%	14%
ATT TMLE	48%	22%	75%	24%	51%	77%	72%	67%

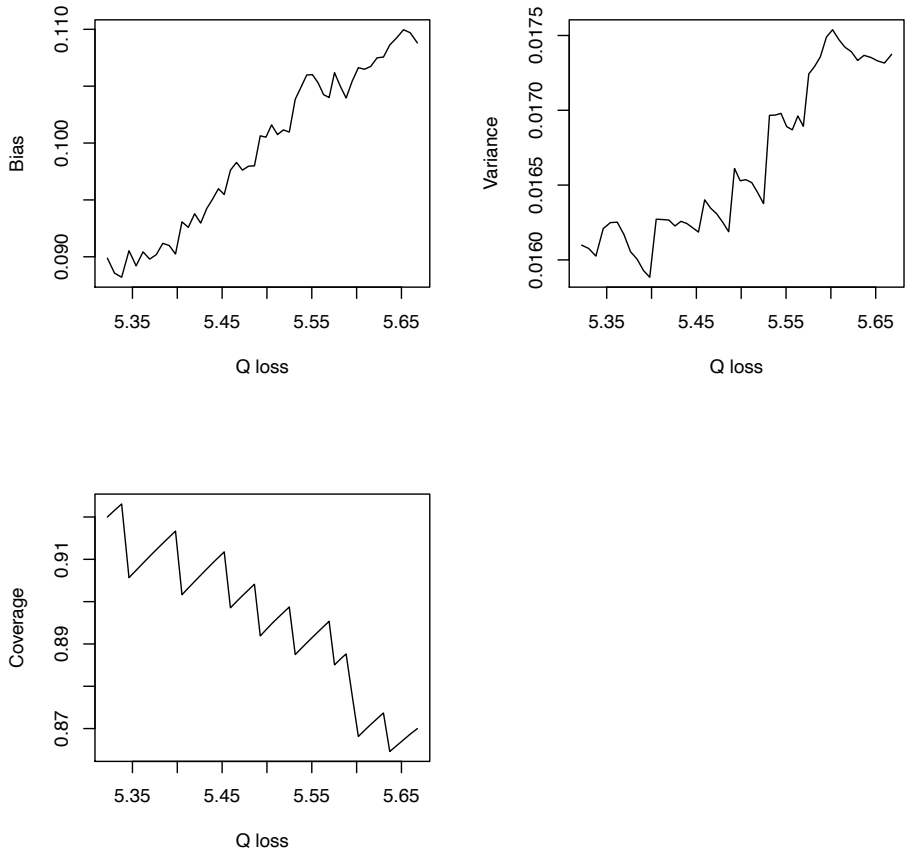

Figure 4: Biases and variances increase while coverages decrease as the mean squared errors of Q (Q loss) becomes larger. This experiment uses 100 datasets with $\beta_t = 1$ (with causal effect), $\beta_c = 50.0$ (low confounding), and $\gamma = 4.0$ (high noise).

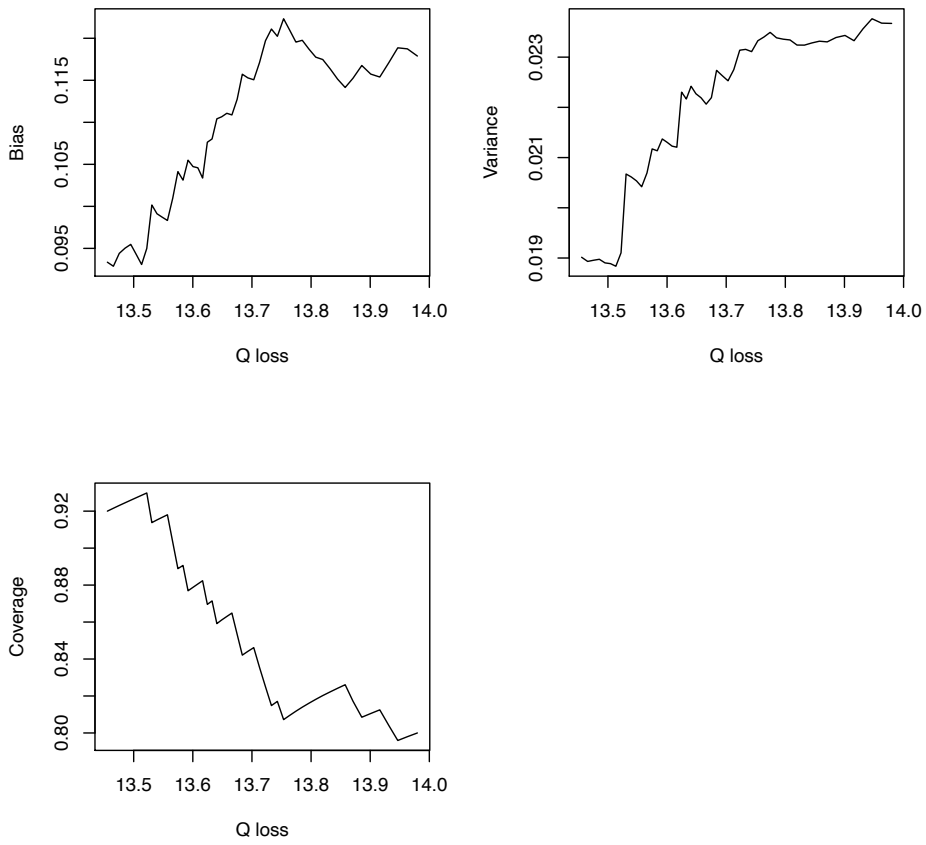

Figure 5: Biases and variances increase while coverages decrease as the mean squared errors of Q becomes larger. This experiment uses 100 datasets with $\beta_t = 1$ (with causal effect), $\beta_c = 100.0$ (high confounding), and $\gamma = 4.0$ (high noise).