Biometrika (2012), **99**, 1, *pp.* 1–15

Printed in Great Britain

Advance Access publication on 5 February 2014

Supplemental Material for Interactive Model Building for Q-Learning

BY ERIC B. LABER, KRISTIN A. LINN AND LEONARD A. STEFANSKI

Department of Statistics, North Carolina State University, 2311 Stinson Drive, 5216 SAS Hall, Raleigh, North Carolina, 27695-8203, USA

laber@stat.ncsu.edu kalinn@ncsu.edu stefansk@ncsu.edu

1. INCONSISTENCY OF Q-LEARNING

The closed-form expression in (10) of the main paper facilitates study of nonlinearity introduced by the nonsmooth maximization operator and resulting inconsistency of Q-learning. To this end, suppose $Q_2(h_2, a_2) = h_{2,0}^{\mathrm{T}}\beta_{2,0}^* + a_2 h_{2,1}^{\mathrm{T}}\beta_{2,1}^*$ so that the second-stage Q-function is correctly specified, and thus $\max_{a_2\in\{-1,1\}} Q_2(H_2, a_2) = H_{2,0}^{\mathrm{T}}\beta_{2,0}^* + |H_{2,1}^{\mathrm{T}}\beta_{2,1}^*|$. Consider the coefficient indexing the best-fitting linear model to the first-stage Q-function,

$$\beta_1^* = \arg\min_{\beta_{1,0},\beta_{1,1}} E\left\{\max_{a_2\in\{-1,1\}} Q_2(H_2, a_2) - H_{1,0}^{\mathrm{T}}\beta_{1,0} - A_1 H_{1,1}^{\mathrm{T}}\beta_{1,1}\right\}^2,$$

so that $\beta_1^* = \Sigma_1^{-1} E B_1\left(H_{2,0}^{\mathrm{T}}\beta_{2,0}^* + |H_{2,1}^{\mathrm{T}}\beta_{2,1}^*|\right)$, where $B_1 = (H_{1,0}^{\mathrm{T}}, A_1 H_{1,1}^{\mathrm{T}})^{\mathrm{T}}$ and $\Sigma_1 = EB_1B_1^{\mathrm{T}}$. If $\mu(H_2) = B_1^{\mathrm{T}}\gamma + \rho$ where ρ is a mean zero random variable which is independent of patient histories and outcomes, then $\beta_1^* = \gamma + \Sigma_1^{-1}EB_1|H_{2,1}^{\mathrm{T}}\beta_{2,1}^*|$, and for $b_1 = (h_{1,0}^{\mathrm{T}}, a_1 h_{1,1}^{\mathrm{T}})^{\mathrm{T}}$ it follows that $b_1^{\mathrm{T}}\beta_1^* = E\{\mu(H_2) \mid h_1, a_1\} + b_1^{\mathrm{T}}\Sigma_1^{-1}EB_1|H_{2,1}^{\mathrm{T}}\beta_{2,1}^*|$. In addition, suppose that $H_{2,1}^{\mathrm{T}}\beta_{2,1}^* = B_1^{\mathrm{T}}\eta + \nu$, where ν is a mean zero normal random variable with variance σ^2 that is independent of patient histories and outcomes. Then $b_1^{\mathrm{T}}\beta_1^* - E\{\mu(H_2)| \mid h_1, a_1\}$ is equal to

$$b_1^{\mathrm{T}}\Sigma^{-1}EB_1\left[B_1^{\mathrm{T}}\eta\left\{1 - 2\Phi\left(\frac{-B_1^{\mathrm{T}}\eta}{\sigma}\right)\right\} + \left(\frac{2\sigma^2}{\pi}\right)^{\frac{1}{2}}\exp\left\{\frac{-(B_1^{\mathrm{T}}\eta)^2}{2\sigma^2}\right\}\right],$$

which can be reexpressed as $E\left(|H_{2,1}^{\mathrm{T}}\beta_{2,1}^*| \mid H_1 = h_1, A_1 = a_1\right) + r(h_1, a_1)$, where

$$\begin{aligned} r(h_1, a_1) = 2b_1^{\mathrm{T}}\Sigma_1^{-1}EB_1B_1^{\mathrm{T}}\eta&\left\{\Phi\left(\frac{-b_1^{\mathrm{T}}\eta}{\sigma}\right) - \Phi\left(\frac{-B_1^{\mathrm{T}}\eta}{\sigma}\right)\right\} \\ &+ \left(\frac{2\sigma^2}{\pi}\right)^{\frac{1}{2}}\left[b_1^{\mathrm{T}}\Sigma^{-1}EB_1\exp\left\{\frac{-(B_1^{\mathrm{T}}\eta)^2}{2\sigma^2}\right\} - \exp\left\{\frac{-(b_1^{\mathrm{T}}\eta)^2}{2\sigma^2}\right\}\right] \end{aligned}$$

is a remainder term that shows how far the optimal linear approximation is from the truth, $E\left(|H_{2,1}^{\mathrm{T}}\beta_{2,1}^*| \mid H_1 = h_1, A_1 = a_1\right) = E\{|\Delta(H_2)| \mid H_1 = h_1, A_1 = a_1\}$. The remainder is identically zero if $\eta = 0$ and $\sigma = 0$, the case of no second-stage treatment effect with probability one, i.e., $\mathrm{pr}(H_{2,1}^{\mathrm{T}}\beta_{2,1}^* = 0) = 1$. The remainder is close to zero when the distribution of $B_1^{\mathrm{T}}\eta/\sigma$ is concentrated sufficiently far from zero and $b_1^{\mathrm{T}}\eta/\sigma$ is also far from zero. The remainder term is largest for small to moderate values of $B_1^{\mathrm{T}}\eta/\sigma$. This is relevant, as in many applications we do not expect large signal-to-noise ratios. Thus, even under simple generative models like the one described above, the Q-learning algorithm with its linear approximations need not be even approximately consistent.

2. PROOFS OF ASYMPTOTIC RESULTS

Let $l^{\infty}(\mathcal{F})$ denote the space of uniformly bounded real-valued functions on $\mathcal{F}$ equipped with the supremum norm. Write $Z_n = n^{1/2}\left(\Delta_L,\ \Delta_m,\ \Delta_\sigma\right)^{\mathrm{T}}$. Then by (A1N), Z_n converges in distribution to $\mathrm{N}\{0,\ \Sigma_N(h_1,a_1)\}$. Similarly, define $W_n = n^{1/2}\left(\Delta_L,\ \Delta_\theta,\ \Delta_\gamma,\ \Delta_\beta,\ \Delta_\xi\right)$. Then by (A1E), W_n converges in distribution to $\mathrm{N}\{0, \Sigma_E(h_1,a_1)\}$. For convenience we abbreviate $m(h_1,a_1;\theta)$, $\sigma(h_1,a_1;\gamma)$, $L(h_1,a_1;\alpha)$, and $\xi(H_2,H_1,A_1;\theta,\gamma,\beta_2)$ as m, σ, L, and ξ, respectively. Similarly, we write $\widehat{m}$, $\widehat{\sigma}$, $\widehat{L}$, and $\widehat{\xi}$ as shorthand for $m(h_1,a_1;\widehat{\theta})$, $\sigma(h_1,a_1;\widehat{\gamma})$, $L(h_1,a_1;\widehat{\alpha})$, and $\xi(H_2,H_1,A_1;\widehat{\theta},\widehat{\gamma},\widehat{\beta}_2)$, and we write m^*, σ^*, L^*, and ξ^* as shorthand for $m(h_1,a_1;\theta^*)$, $\sigma(h_1,a_1;\gamma^*)$, $L(h_1,a_1;\alpha^*)$, and $\xi(H_2,H_1,A_1;\theta^*,\gamma^*,\beta_2^*)$.

Proof of Theorem 1, Part 1. Notice that

$$n^{1/2}\left\{\widehat{Q}_1^{\mathrm{IQ},N}(h_1,a_1) - L(h_1,a_1;\alpha^*) - \frac{1}{\sigma^*}\int |z|\phi\left(\frac{z-m^*}{\sigma^*}\right)dz\right\} \\ = n^{1/2}\left\{I(\widehat{L},\widehat{m},\widehat{\sigma}) - I(L^*,m^*,\sigma^*)\right\},$$

where $I(\cdot)$ is as defined immediately preceding Theorem 1 in the main paper. Inspection reveals that $\nabla I(L,m,\sigma)$ exists and is continuous in a neighborhood of (L^*,m^*,σ^*). Hence, by a first-order Taylor series approximation, the right hand side above is equal to

$$n^{1/2}\left\{I(\widehat{L},\widehat{m},\widehat{\sigma}) - I(L^*,m^*,\sigma^*)\right\} = \nabla I(L^*,m^*,\sigma^*)^{\mathrm{T}} Z_n + o_P(1).$$

The result follows from Slutsky's lemma. ■

Remark 1. It is possible to extend the above proof to obtain bootstrap consistency. Let $E^{(b)}$ denote the bootstrap empirical distribution. We use $u^{(b)}$ to denote the bootstrap analog of functional u, e.g., $u = u(E_n, E)$ then $u^{(b)} = u(E_n^{(b)}, E_n)$. If, in addition to the conditions for Theorem 1, $Z_n^{(b)}$ converges weakly in probability to $N\{0, \Sigma_N(h_1,a_1)\}$, then the above proof goes through using exactly the same arguments after changing $I(\widehat{L},\widehat{m},\widehat{\sigma})$ to $I(\widehat{L}^{(b)},\widehat{m}^{(b)},\widehat{\sigma}^{(b)})$ and $I(L^*,m^*,\sigma^*)$ to $I(\widehat{L},\widehat{m},\widehat{\sigma})$ (see Kosorok, 2008 for bootstrap continuous mapping theorems and bootstrap central limit theorems).

Proof of Theorem 1, Part 2. The proof proceeds by showing that

$$n^{1/2}\left\{\widehat{Q}_1^{\mathrm{IQ},E}(h_1,a_1) - L(h_1,a_1;\alpha^*) - \frac{1}{\sigma^*}\int |z|\kappa\left(\frac{z-m^*}{\sigma^*}\right)dz\right\} \\ = \{1, \nabla J(\theta^*,\gamma^*,\beta_2^*)^{\mathrm{T}}, 1\}W_n + o_P(1). \quad (1)$$

The term on the left hand side of the above display equals

$$n^{1/2}E_n|\widehat{m} + \widehat{\sigma}\widehat{\xi}| - n^{1/2}E|m^* + \sigma^*\xi^*| + W_{n,1}.$$

The first two terms in the above display are equal to

$$n^{1/2}(E_n - E)|\widehat{m} + \widehat{\sigma}\widehat{\xi}| + n^{1/2}E\left(|\widehat{m} + \widehat{\sigma}\widehat{\xi}| - |m^* + \sigma^*\xi^*|\right).$$

From (A2), it follows that $n^{1/2}(E_n - E)$ converges weakly to G_∞ in $l^\infty(\mathcal{F})$, where G_∞ is a mean zero Gaussian process with covariance function $\mathrm{Cov}\{G_\infty(f), G_\infty(g)\} = E(f - Ef)(g - Eg)$ (see, for example, Kosorok, 2008). Note that by the second part of (A2), the foregoing covariance function is continuous in a neighborhood of $(\theta^*,\gamma^*,\beta_2^*)$. Thus, using the equicontinuity of $n^{1/2}(E_n - E)$, it follows that $n^{1/2}(E_n - E)|\widehat{m} + \widehat{\sigma}\widehat{\xi}| = W_{n,5} + o_{P^*}(1)$, where P^* denotes outer probability. So far, we have shown that the right hand side of (1) is equal to $n^{1/2}\left\{E(|\widehat{m} + \widehat{\sigma}\widehat{\xi}| - |m^* + \sigma^*\xi^*|)\right\} + W_{n,1} + W_{n,5} + o_{P^*}(1)$. From (A2), J is continuously differentiable in a neighborhood of $(\theta^*, \gamma^*, \beta_2^*)$. Using a first-order Taylor series approximation, we have

$$n^{1/2}E(|\widehat{m} + \widehat{\sigma}\widehat{\xi}| - |m^* + \sigma^*\xi^*|) = \nabla J(\theta^*, \gamma^*, \beta_2^*)^{\mathrm{T}}(W_{n,2}, W_{n,3}, W_{n,4}) + o_P(1).$$

Thus, we have shown that the right hand side of (1) equals $\{1, \nabla J(\theta^*, \gamma^*, \beta_2^*)^{\mathrm{T}}, 1\}W_n + o_{P^*}(1)$. The result follows from Slutsky's Lemma (Kosorok, 2008). ■

Remark 2. It is possible under mild conditions to extend the above proof to obtain bootstrap consistency, e.g., that $W^{(b)}$ converges weakly in probability to $N\{0, \Sigma_E(h_1, a_1)\}$. Note, for example, that the bootstrap empirical process $n^{1/2}(E_n^{(b)} - E_n)$ converges weakly in probability to G_∞ in $l^\infty(\mathcal{F})$ by (A2) and Theorem 2.6 in Kosorok (2008).

3. Obtaining asymptotic normality of IQ-learning parameters

Here we provide a sketch of how one obtains asymptotic normality of the parameters used in IQ-learning. For a more complete discussion of conditional variance estimators and proofs of asymptotic normality under more general conditions see Carroll and Ruppert (1988). For illustration we use the working models from the simulated experiments in the main body. We demonstrate using $\widehat{\gamma}$ as this is the most involved; other estimators would be handled similarly. We assume linear models for $\Delta(H_2)$ and $m(H_1, A_1)$ so that $\Delta(H_2; \beta_2) = H_{2,1}^{\mathrm{T}}\beta_{2,1}$ and $m(H_1, A_1; \theta) = H_{1,0}^{\mathrm{T}}\theta_{1,0} + A_1 H_{1,1}^{\mathrm{T}}\theta_{1,1}$. We assume a log-linear model for $\sigma(H_1, A_1)$ so that $\log\sigma(H_1, A_1; \gamma) = H_{1,0}^{\mathrm{T}}\gamma_{1,0} + A_1 H_{1,1}^{\mathrm{T}}\gamma_{1,1}$. Define $G = (H_{2,1}^{\mathrm{T}}, -H_{1,0}, -A_1 H_{1,1}^{\mathrm{T}})^{\mathrm{T}}$, $\widehat{\Gamma} = (\widehat{\beta}_{2,1}^{\mathrm{T}}, \widehat{\theta}^{\mathrm{T}})^{\mathrm{T}}$, and $\Gamma^* = (\beta_{2,1}^{*\mathrm{T}}, \theta^{*\mathrm{T}})^{\mathrm{T}}$. We assume that $n^{1/2}(\widehat{\Gamma} - \Gamma^*) = n^{1/2}(E_n - E)s(H_2, H_1, A_1; \Gamma^*) + o_P(1)$ for square integrable score function s. Define $B_1 = (H_{1,0}^{\mathrm{T}}, A_1 H_{1,1}^{\mathrm{T}})^{\mathrm{T}}$. We also assume that $E||B_1||\,||G||\,|G^{\mathrm{T}}\Gamma|^{-1} < \infty$ for all Γ in a neighborhood of Γ^*. Then,

$$\widehat{\gamma} = \arg\min_{\gamma} E_n\left(\log|G^{\mathrm{T}}\widehat{\Gamma}| - B_1^{\mathrm{T}}\gamma\right)^2.$$

Differentiating and setting to zero yields

$$\widehat{\gamma} = \left(E_n B_1 B_1^{\mathrm{T}}\right)^{-1} E_n B_1 \log|G^{\mathrm{T}}\widehat{\Gamma}|.$$

Add and subtract γ^* to the above equality and scale by $n^{1/2}$ to obtain

$$n^{1/2}(\widehat{\gamma} - \gamma^*) = \left(E_n B_1 B_1^{\mathrm{T}}\right)^{-1} n^{1/2} E_n B_1\left(\log|G^{\mathrm{T}}\widehat{\gamma}| - B_1^{\mathrm{T}}\gamma^*\right).$$

After some algebra, it can seen that $n^{1/2}(\widehat{\gamma} - \gamma^*)$ is equal to

$$\left(E_n B_1 B_1^{\mathrm{T}}\right)^{-1} n^{1/2}(E_n - E)B_1\left(\log|G^{\mathrm{T}}\Gamma^*| - B_1^{\mathrm{T}}\gamma^*\right) + \left(E_n B_1 B_1^{\mathrm{T}}\right)^{-1} n^{1/2} E_n B_1\left(\log|G^{\mathrm{T}}\widehat{\Gamma}| - \log|G^{\mathrm{T}}\Gamma^*|\right). \quad (2)$$

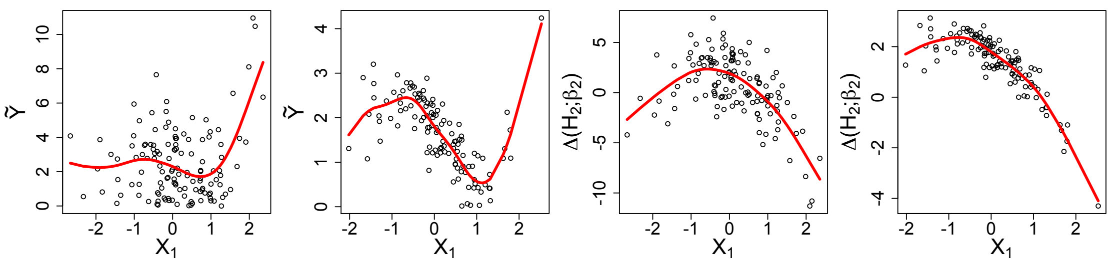

Fig. 1: Detecting quadratic relationships at stage one. From the left, the first two panels are scatterplots of $\widetilde{Y}$ against X_1 for $A_1 = 1$ and $A_1 = -1$, respectively; the true quadratic relationship is masked by the nonsmooth transformation of data. The third and forth panels contain scatterplots of the contrast $\Delta(H_2; \beta_2)$ against X_1 by treatment $A_1 = 1$ and $A_1 = -1$, respectively; the true quadratic relationship is clearly distinguishable. Red lines are cubic smoothing spline fits to the data.

Let $\widetilde{\Gamma}$ be intermediate to Γ^* and $\widehat{\Gamma}$. Then, a Taylor series expansion applied to the second term of (2) shows $n^{1/2}(\widehat{\gamma} - \gamma)$ is equal to

$$
\begin{aligned}
&\left(E_n B_1 B_1^{\mathrm{T}}\right)^{-1} n^{1/2}(E_n - E)\left(\log|G^{\mathrm{T}}\Gamma^*| - B_1^{\mathrm{T}}\gamma^*\right) \\
&\qquad\qquad + \left(E_n B_1 B_1^{\mathrm{T}}\right)^{-1} E_n B_1 (G^{\mathrm{T}}\widetilde{\Gamma})^{-1} G^{\mathrm{T}} n^{1/2}(\widehat{\Gamma} - \Gamma^*) \\
&= \left(E B_1 B_1^{\mathrm{T}}\right)^{-1} n^{1/2}(E_n - E)\left[\log|G^{\mathrm{T}}\Gamma^*| - B_1^{\mathrm{T}}\gamma^* + E\left\{B_1 (G^{\mathrm{T}}\Gamma^*)^{-1} G^{\mathrm{T}}\right\} s(H_1, A_1, H_2; \Gamma^*)\right] \\
&\qquad\qquad + o_P(1),
\end{aligned}
$$

which is asymptotically normal by the central limit theorem and Slutsky's theorem.

4. POWER TO DETECT A QUADRATIC EFFECT

One strength of IQ-learning is that it enables practitioners to apply standard interactive model building techniques. We now consider a generative model with a univariate predictor X_1 and nonlinear relationship between X_1 and X_2. The new generative model is

$$
\begin{aligned}
X_1 &\sim \mathrm{Normal}(.1, 1),\ A_t \sim \mathrm{Uniform}\{-1, 1\},\ t = 1, 2, \\
X_2 &= X_1^2 + (1.5 - 0.5A_1)X_1 + \zeta_{A_1}\xi,\ \xi \sim \mathrm{Normal}(0, 1), \\
\phi &\sim \mathrm{Normal}(0, 4),\ Y = H_{2,0}^{\mathrm{T}}\beta_{2,0} + A_2 H_{2,1}^{\mathrm{T}}\beta_{2,1} + \phi,
\end{aligned}
$$

where $H_{2,0} = H_{2,1} = (1, X_2, A_1, A_1X_2)^{\mathrm{T}}$ and $\zeta_{A_1} = (1.5 + 0.5A_1)^{1/2}$. Thus, the true first-stage Q-function depends on both X_1^2 and $A_1X_1^2$. As in the main paper, we fix $\beta_{2,0}$ and scale $\beta_{2,1}$. We specify the second-stage as

$$
\beta_{2,0} = \frac{(3, -1.5, .4, -1)^{\mathrm{T}}}{||(3, -1.5, .4, -1)^{\mathrm{T}}||},\ \ \beta_{2,1} = C\frac{(2, -1, .2, -.5)^{\mathrm{T}}}{||(2, -1, .2, -.5)^{\mathrm{T}}||},
$$

for $C \in (0, 2)$. Figure 1 illustrates how the quadratic effect of X_1 is masked by the absolute value operator in Q-learning. Alternatively, the quadratic relationship is clearly visible in the scatter plots of the contrast function $\Delta(H_2; \beta_2)$ against X_1. The solid lines in Figure 1 are cubic smoothing splines fitted to the data using ordinary cross validation.

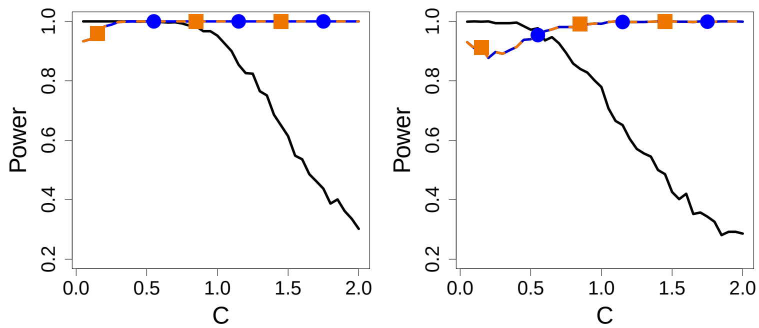

Fig. 2: Power to detect X_1^2 (left) and $A_1X_1^2$ (right). Blue lines with circles, orange dashed lines with squares, and black solid lines represent the normal IQ-learning estimator, nonparametric IQ-learning estimator, and Q-learning, respectively.

Figure 2 displays plots of the power to detect the quadratic effects X_1^2 and $A_1X_1^2$ as a function of the second-stage effect size scaling constant C. The Q-learning curve represents the power to detect the quadratic terms in the regression of the pseudo outcome $\tilde{Y}$ on the first-stage history and treatment. The two identical IQ-learning curves represent the power to detect the quadratic effects in the regression of the contrast function on the first-stage information, that is, the fit of the contrast function mean. Results are based on $n = 250$ training samples, and the power was calculated by averaging over indicators from M =1,000 Monte Carlo data sets of whether the estimated coefficients of X_1^2 and $A_1X_1^2$ were found to be significant by a t-test. When the treatment interaction effects are near zero, i.e., $C \approx 0$, Q-learning detects the nonlinear relationships because the pseudo outcome $\tilde{Y}$ is dominated by the linear main-effect term $H_{2,0}^{\mathrm{T}}\beta_{2,0}$, which is a function of both X_1^2 and $A_1X_1^2$. At first glance, IQ-learning appears to perform worse than Q-learning when the effect size is small. However, this is due to the fact that Figure 2 only displays results from the regression of the contrast function on first-stage information, and $C \approx 0$ implies $\Delta(H_2; \beta_2) \approx 0$. Results from the regression of the main-effect term $H_{2,0}^{\mathrm{T}}\beta_{2,0}$ on first-stage information are not included in Figure 2.

The power of Q-learning to detect the quadratic terms decreases drastically as the second-stage treatment effects increase because the absolute value from the maximization operator masks the true underlying structure. We note that the parameters that index the first-stage Q-function are nonregular, so the t-tests for significance are invalid. In comparison, the first-stage IQ-learning coefficients are asymptotically normal. Thus t-tests are approximately valid and they detect the quadratic relationships in the mean of the contrast function with increasing accuracy as the treatment effects grows larger.

In Figures 3 and 4, we provide results from the same model when all first-stage IQ- and Q-learning models include linear terms only and all first-stage IQ- and Q-learning models include a quadratic term, respectively. Although linear Q-learning outperforms both misspecified linear IQ-learning estimators in terms of integrated mean squared error, the average value of the nonparametric IQ-learning estimator is comparable with Q-learning. In addition, the correctly specified quadratic version of the nonparametric IQ-learning outperforms Q-learning with quadratic terms with respect to all four displayed measures of performance.

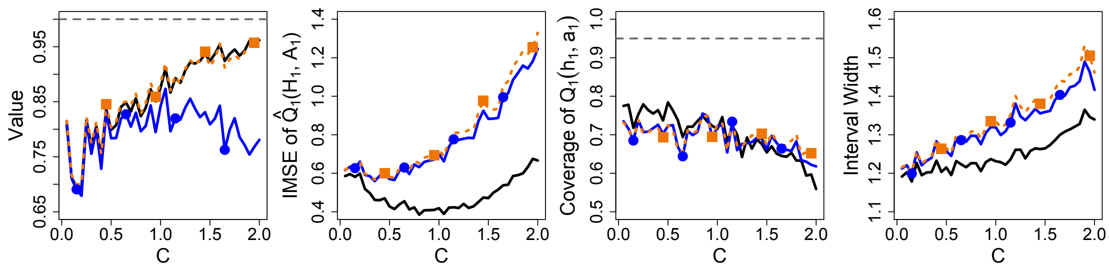

Fig. 3: Results for IQ-learning and Q-learning with linear first-stage model terms only. Blue lines with circles, orange dashed lines with squares, and black solid lines represent the normal IQ-learning estimator, nonparametric IQ-learning estimator, and Q-learning, respectively.

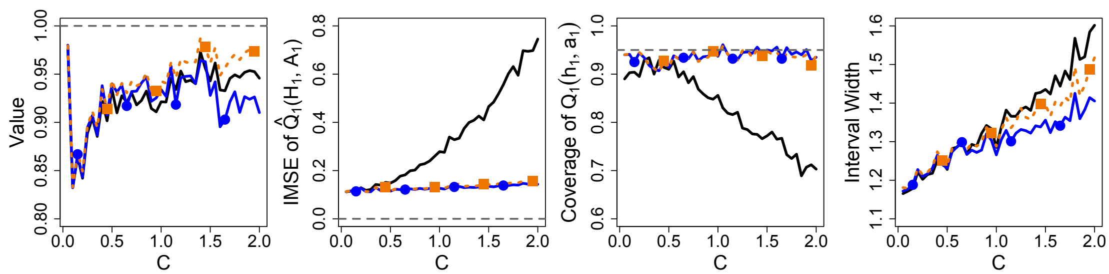

Fig. 4: Results for IQ-learning and Q-learning with quadratic terms included in all first-stage models. Blue lines with circles, orange dashed lines with squares, and black solid lines represent the normal IQ-learning estimator, nonparametric IQ-learning estimator, and Q-learning, respectively.

5. ADDITIONAL SIMULATION RESULTS

Here we provide additional simulation results to demonstrate the robust performance of IQ-learning across a broad range of model settings. As in the main portion of the paper, the generative model is

$$X_1 \sim \text{Normal}_p\{0.1, \Omega_{AR_1}(0.5)\}, \;\; A_t \sim \text{Uniform}\{-1, 1\}, \, t = 1, 2,$$
$$X_2 = (1.5 - 0.5A_1)X_1 + \zeta_{A_1}\xi, \;\; Y = H_2^{\mathrm{T}}\beta_{2,0} + A_2 H_2^{\mathrm{T}}\beta_{2,1} + \phi,$$

where $\{\Omega_{\mathrm{AR}_1}(0.5)\}_{i,j} = (0.5)^{|i-j|}$, $H_2 = (1, X_2^{\mathrm{T}}, A_1, A_1 X_2^{\mathrm{T}})^{\mathrm{T}}$, and $\zeta_{A_1} = (1.5 + 0.5A_1)^{1/2}$. Thus, the class is indexed by the dimension p, the distributions of ξ and ϕ, and the coefficient vectors $\beta_{2,0}$ and $\beta_{2,1}$. We fix the main effect parameter $\beta_{2,0}$ and vary the second-stage treatment effect size by scaling $\beta_{2,1}$ as follows:

$$\beta_{2,0} = \frac{1_{2p+2}}{||1_{2p+2}||}, \qquad \beta_{2,1} = C\frac{(-0.25 \cdot 1_{p+1}^{\mathrm{T}}, 1_{p+1}^{\mathrm{T}})^{\mathrm{T}}}{||(-0.25 \cdot 1_{p+1}^{\mathrm{T}}, 1_{p+1}^{\mathrm{T}})||},$$

where C ranges over a grid from 0 to 2, and 1_d denotes a d-dimensional vector of 1s. In addition, we fix the theoretical R^2 of the second-stage regression model by generating $\phi \sim \text{Normal}\{0, \sigma_\phi^2(C)\}$, where the variance $\sigma_\phi^2(C)$ depends on the scaling constant C. We consider training sets of size $n = 250$ and $n = 500$ and vary the second-stage $R^2 \in \{0.4, 0.6, 0.8\}$. Re-

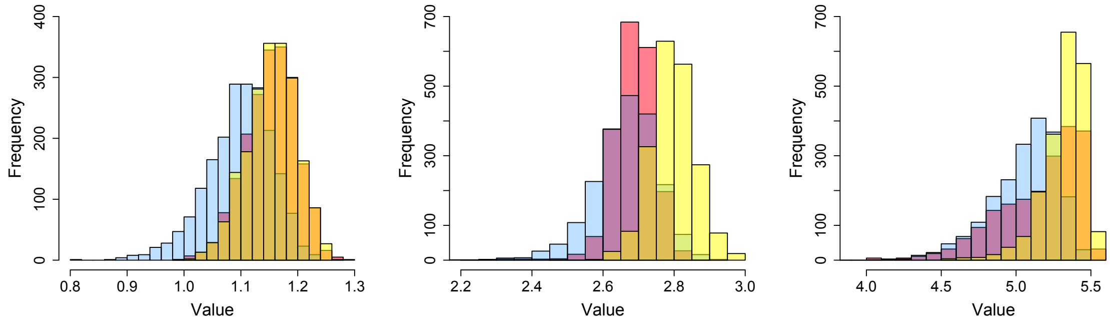

Fig. 5: Histograms of the value estimates from each Monte Carlo iteration for, left to right, C=0.05, 1.0, 2.0. Results from Q-learning with linear models, Q-learning with Support Vector Regression, and `NormHomo` IQ-learning are shown in red, blue, and yellow, respectively.

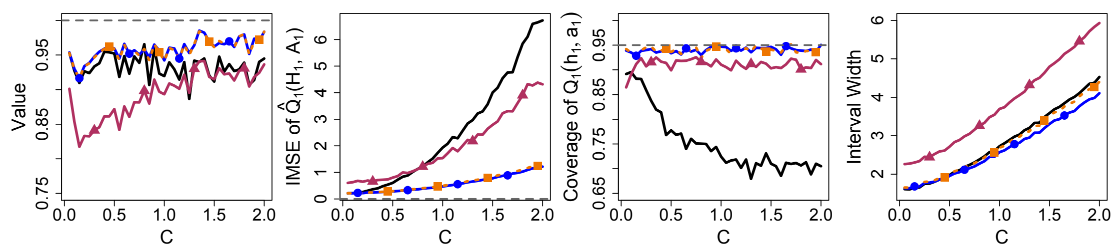

Fig. 6: Measures of performance of the normal IQ-learning estimator, nonparametric IQ-learning estimator, Q-learning with linear models, and support vector regression Q-learning represented by blue lines with circles, orange dashed lines with squares, black solid lines, and maroon solid lines with triangles, respectively; elements of ξ generated independently from t_5; $R^2 = 0.6$; p = 4; $n = 250$. From left to right: average proportion of optimal value obtained; integrated mean squared error of Q_1 estimates; coverage of 95% confidence intervals for Q_1; width of 95% confidence intervals for Q_1.

sults for $n = 250$ with $R^2 = 0.6$ are included in Section 3 of the paper. In this section, we provide results for the remaining combinations of n and R^2. We include simulations with ξ generated from a $\text{Normal}_p(0, I_p)$ as well as where elements of ξ generated independently from a t-distribution with five degrees of freedom. We include results for dimension $p = 4$, followed by results for $p = 8$ when $R^2 = 0.6$. In each simulation, results are based on $M = 2,000$ Monte Carlo data sets.

Figure 5 displays additional results regarding the value, $V^\pi = E^\pi Y$, of the estimated regimes from Section 3 of the main paper. Histograms of the value estimates from each Monte Carlo iteration from the normal IQ-learning estimator, Q-learning with linear models, and support vector regression Q-learning are displayed in Figure 5 for three values of the scaling constant, C. In general, the estimated value distribution of the IQ-learning estimated regime is shifted slightly higher than both of the Q-learning estimated value distributions. Results shown are for $C = 0.05, 1.0, 2.0$; results were similar across other values of C.

Figure 6 presents results when $R^2 = 0.6$, $p = 4$, $n = 250$, and elements of ξ generated independently from a t-distribution with five degrees of freedom.

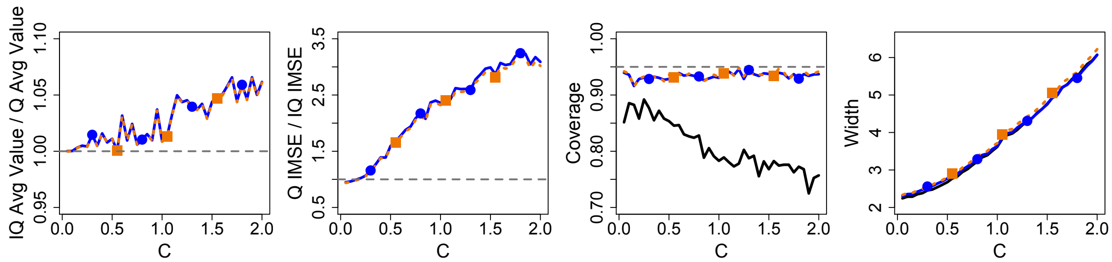

Fig. 7: Measures of performance of the normal IQ-learning estimator, nonparametric IQ-learning estimator and Q-learning represented by blue lines with circles, orange dashed lines with squares, and black solid lines, respectively; $\xi \sim \text{Normal}_p(0, I_p)$; $R^2 = 0.4$; $p = 4$; $n = 250$. From left to right: ratio of average value, coded so values greater than one are favorable to IQ-learning; integrated mean squared error ratio of Q_1 estimates, coded so values greater than one are favorable to IQ-learning; coverage of 95% confidence intervals for Q_1; width of 95% confidence intervals for Q_1.

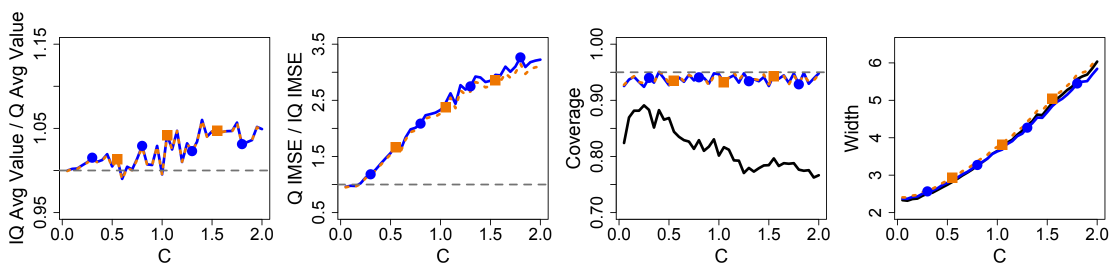

Fig. 8: Measures of performance of Q-learning vs. IQ-learning; components of ξ generated independently from t_5; $R^2 = 0.4$; $p = 4$; $n = 250$.

Define $H_1 = (1, X_1^{\mathrm{T}})^{\mathrm{T}}$. As in Section 3, we consider linear working models for the mean and variance functions of the form

$$Q_2(h_2, a_2; \beta_2) = h_2^{\mathrm{T}}\beta_{2,0} + a_2 h_2^{\mathrm{T}}\beta_{2,1}, \;\; Q_1(h_1, a_1; \beta_1) = h_1^{\mathrm{T}}\beta_{1,0} + a_1 h_1^{\mathrm{T}}\beta_{1,1},$$
$$L(h_1, a_1; \alpha) = h_1^{\mathrm{T}}\alpha_0 + a_1 h_1^{\mathrm{T}}\alpha_1, \;\; m(h_1, a_1; \theta) = h_1^{\mathrm{T}}\theta_0 + a_1 h_1^{\mathrm{T}}\theta_1,$$
$$\log\{\sigma(h_1, a_1; \gamma)\} = h_1^{\mathrm{T}}\gamma_0 + a_1 h_1^{\mathrm{T}}\gamma_1.$$

In Section 3, we considered two IQ-learning estimators: the normal estimator $\widehat{g}^N_{h_1,a_1}(\cdot)$ of the residual distribution and a restricted variance model, $\log\{\sigma(h_1, a_1; \gamma)\} = \gamma_0 + a_1\gamma_1$, that depends only on treatment; and the nonparametric estimator $\widehat{g}^E_{h_1,a_1}(\cdot)$ of the residual distribution with a log-linear variance model that depends on h_1 and a_1. When $\xi \sim \text{Normal}_p(0, I_p)$, both these estimators are correctly specified. When the elements of ξ are generated independently from t_5, only the nonparametric estimator is correctly specified.

Figures 7 - 20 display the results. For all settings, the integrated mean squared error ratio of Q-learning to IQ-learning is greater than one, indicating that the IQ-learning estimators more accurately estimate the first-stage Q-funciton. Increasing the sample size to $n = 500$ and specifying higher R^2 values leads to the greatest gains in integrated mean squared error of IQ-learning compared to Q-learning. In general, coverage of 95% confidence intervals for Q_1 and average value ratios seem consistent across all settings of the parameters. In particular, the IQ-learning

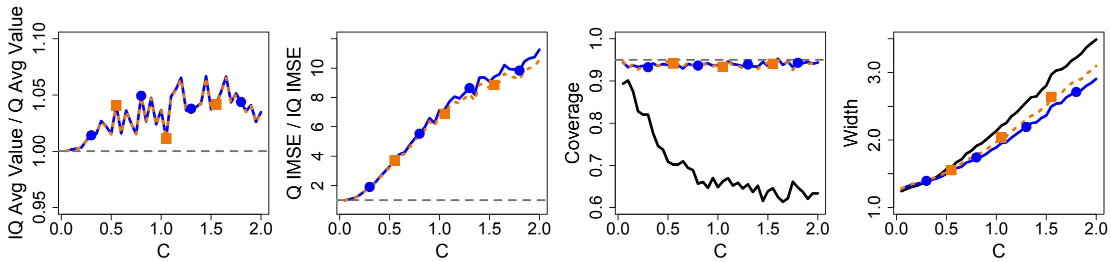

Fig. 9: Measures of performance of Q-learning vs. IQ-learning; $\xi \sim \text{Normal}_p(0, I_p)$; $R^2 = 0.8$; $p = 4$; $n = 250$.

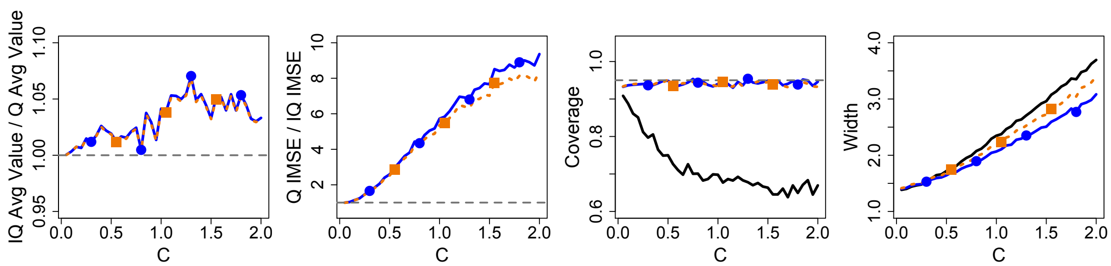

Fig. 10: Measures of performance of Q-learning vs. IQ-learning; components of ξ generated independently from t_5; $R^2 = 0.8$; $p = 4$; $n = 250$.

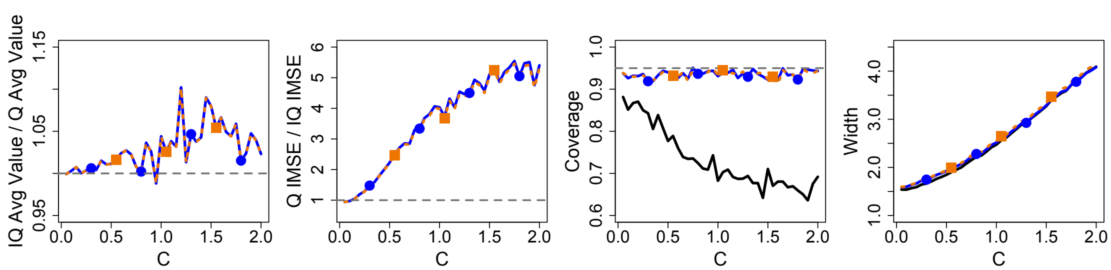

Fig. 11: Measures of performance of Q-learning vs. IQ-learning; $\xi \sim \text{Normal}_p(0, I_p)$; $R^2 = 0.4$; $p = 4$; $n = 500$.

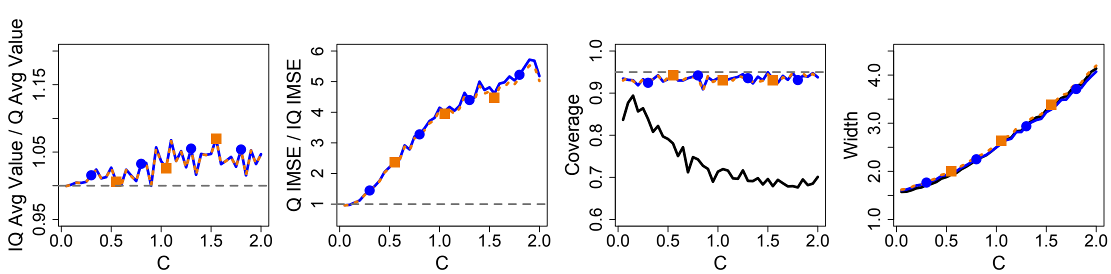

Fig. 12: Measures of performance of Q-learning vs. IQ-learning; components of ξ generated independently from t_5; $R^2 = 0.4$; $p = 4$; $n = 500$.

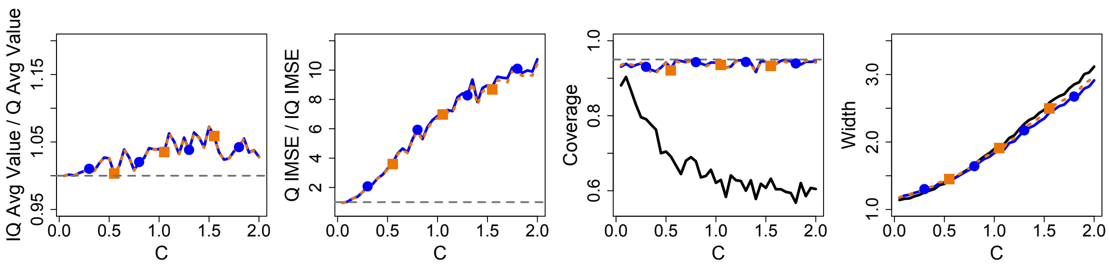

Fig. 13: Measures of performance of Q-learning vs. IQ-learning; $\xi \sim \text{Normal}_p(0, I_p)$; $R^2 = 0.6$; $p = 4$; $n = 500$.

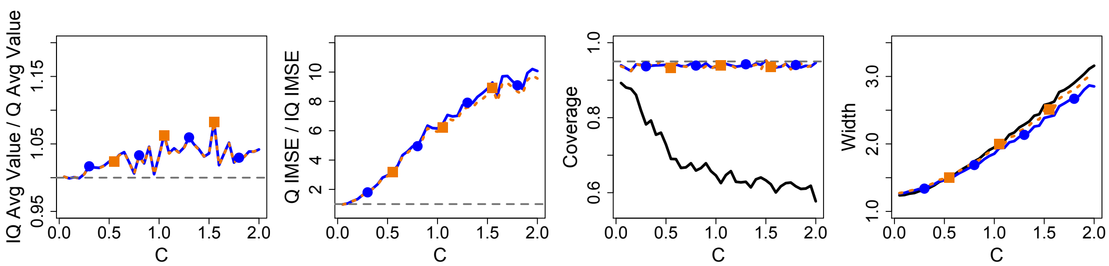

Fig. 14: Measures of performance of Q-learning vs. IQ-learning; components of ξ generated independently from t_5; $R^2 = 0.6$; $p = 4$; $n = 500$.

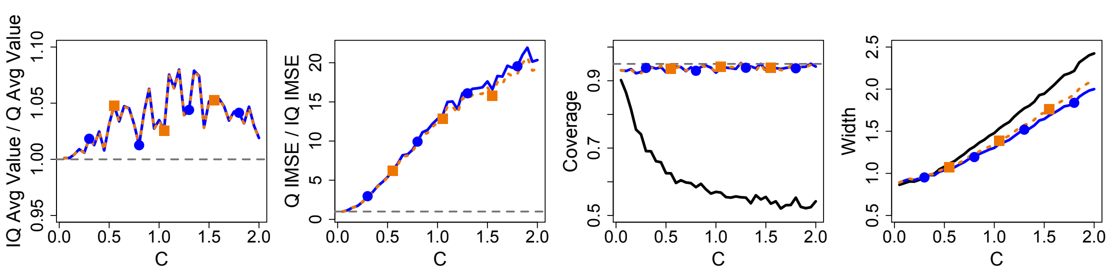

Fig. 15: Measures of performance of Q-learning vs. IQ-learning; $\xi \sim \text{Normal}_p(0, I_p)$; $R^2 = 0.8$; $p = 4$; $n = 500$.

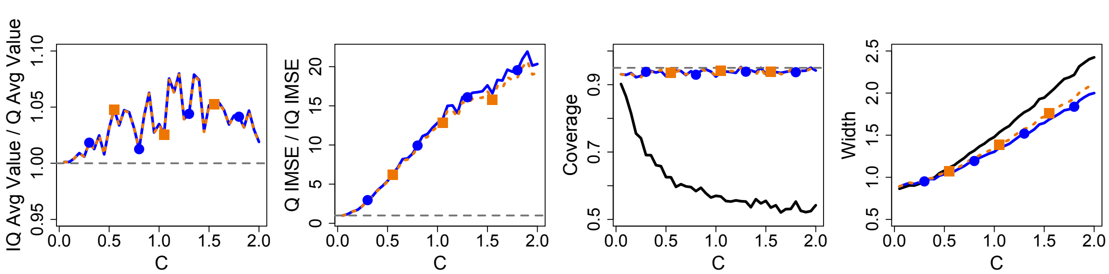

Fig. 16: Measures of performance of Q-learning vs. IQ-learning; components of ξ generated independently from t_5; $R^2 = 0.8$; $p = 4$; $n = 500$.

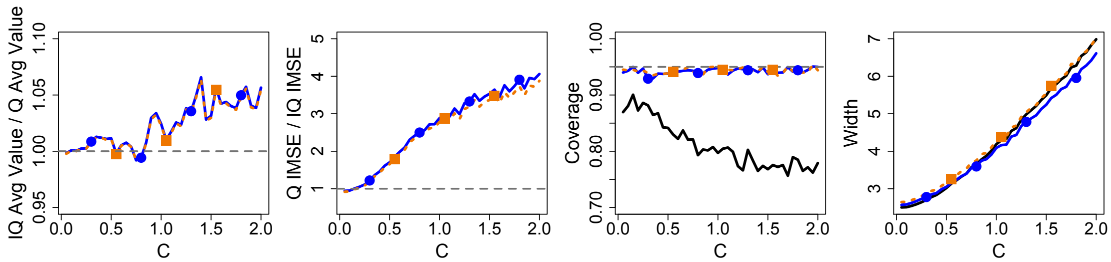

Fig. 17: Measures of performance of Q-learning vs. IQ-learning; $\xi \sim \text{Normal}_p(0, I_p)$; $R^2 = 0.6$; $p = 8$; $n = 250$.

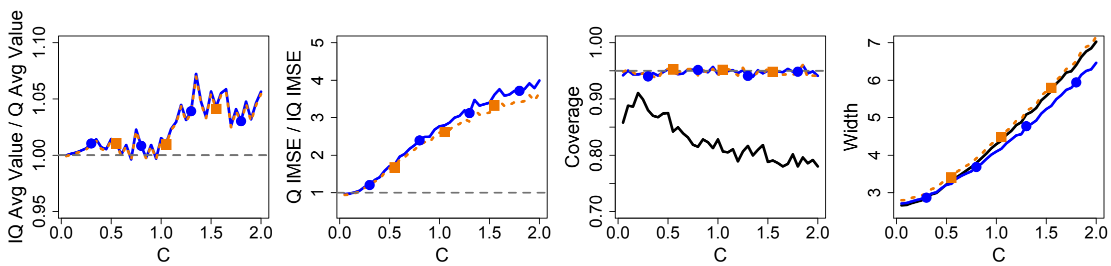

Fig. 18: Measures of performance of Q-learning vs. IQ-learning; components of ξ generated independently from t_5; $R^2 = 0.6$; $p = 8$; $n = 250$.

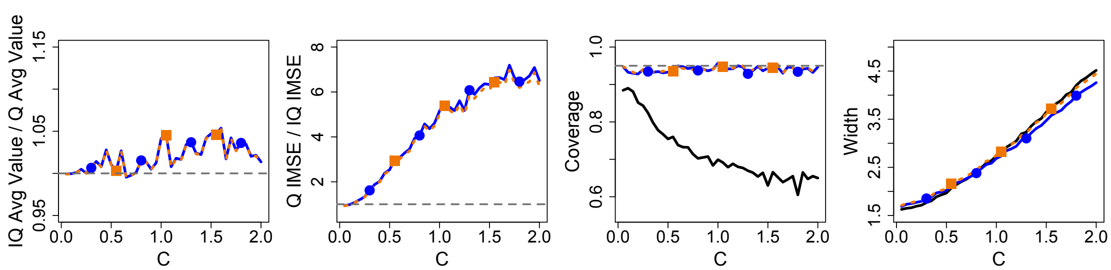

Fig. 19: Measures of performance of Q-learning vs. IQ-learning; $\xi \sim \text{Normal}_p(0, I_p)$; $R^2 = 0.6$; $p = 8$; $n = 500$.

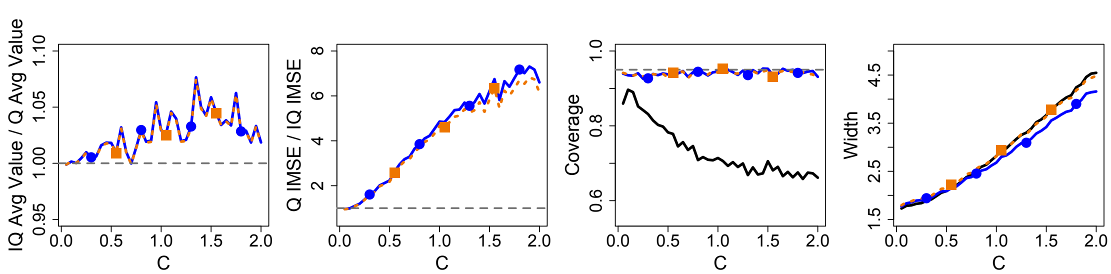

Fig. 20: Measures of performance of Q-learning vs. IQ-learning; components of ξ generated independently from t_5; $R^2 = 0.6$; $p = 8$; $n = 500$.

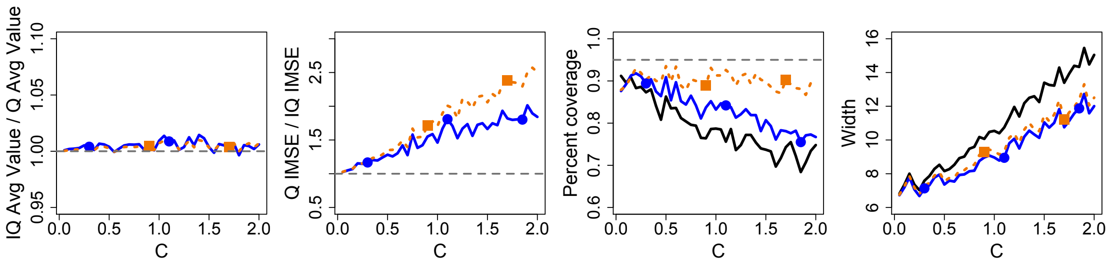

Fig. 21: Measures of performance of Q-learning vs. IQ-learning; components of ξ generated independently from $\mathrm{Lognormal}(0, 1)$;

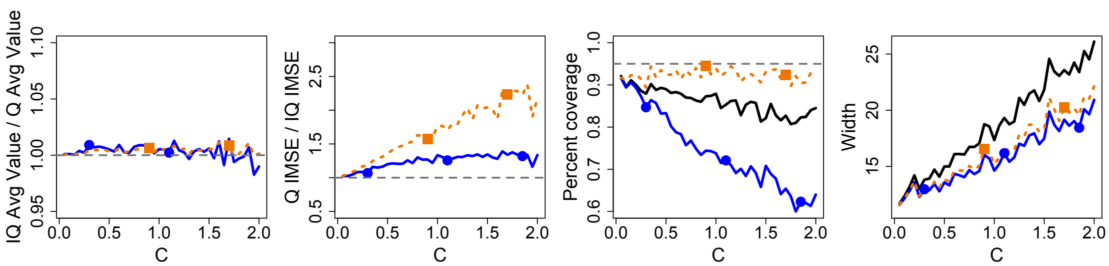

Fig. 22: Measures of performance of Q-learning vs. IQ-learning; components of ξ are independent draws from a mixture of $\mathrm{Normal}(4, 1)$ and $\mathrm{Normal}(0, 1)$, each with probability 0.5.

estimators obtain close to the 95% nominal coverage level in all settings across values of C, while Q-learning suffers from poor coverage, especially for high R^2 values and large effect sizes.

Results in Figures 21 and 22 arise from generative models where only the nonparametric IQ-learning estimator is correctly specified. In these settings, we vary the distribution ξ and substitute ζ_{X_1,A_1} for ζ_{A_1}. That is, we specify a variance model that depends on both the first-stage treatment and first-stage covariates according to the relationship $\zeta_{X_1,A_1} = \exp[\log(2)/4 + .25^{\mathrm{T}}_{p-1}X_1 + A_1\{\log(2)/4 + .1^{\mathrm{T}}_{p-1}X_1\}]$. Results are based on $n = 250$ training samples, $M = 1,000$ Monte Carlo data sets, and dimension $p = 4$. The second-stage R^2 is not fixed. Figure 21 presents results from the case where the components of ξ are generated independently from a Lognormal(0,1) distribution. Figure 22 results arise when elements of ξ are drawn independently from a mixture of the Normal(4,1) and Normal(0,1) distributions. The nonparametric IQ-learning estimator clearly outperforms the normal estimator in Figures 21 and 22, whereas their performance is nearly indistinguishable when both are correctly specified.

In Figure 21, we see that both IQ-learning estimators improve integrated mean squared error over Q-learning, with the nonparametric IQ-learning estimator achieving greater gains in performance. In addition, the coverage plot in Figure 21 shows that the IQ-learning estimators fall short of the nominal 95% level, even though the widths of these confidence intervals are much larger than those observed in the correctly specified simulations. Coverage is still improved when compared to Q-learning. The nonparametric IQ-learning estimator achieves the highest coverage at nearly 90% for most values of C. The ratio of average value is near one for both IQ-learning estimators, indicating little difference in the mean of the final response when treating according to IQ-learning or Q-learning estimated regimes. The results in Figure 22 are similar to those in Fig-

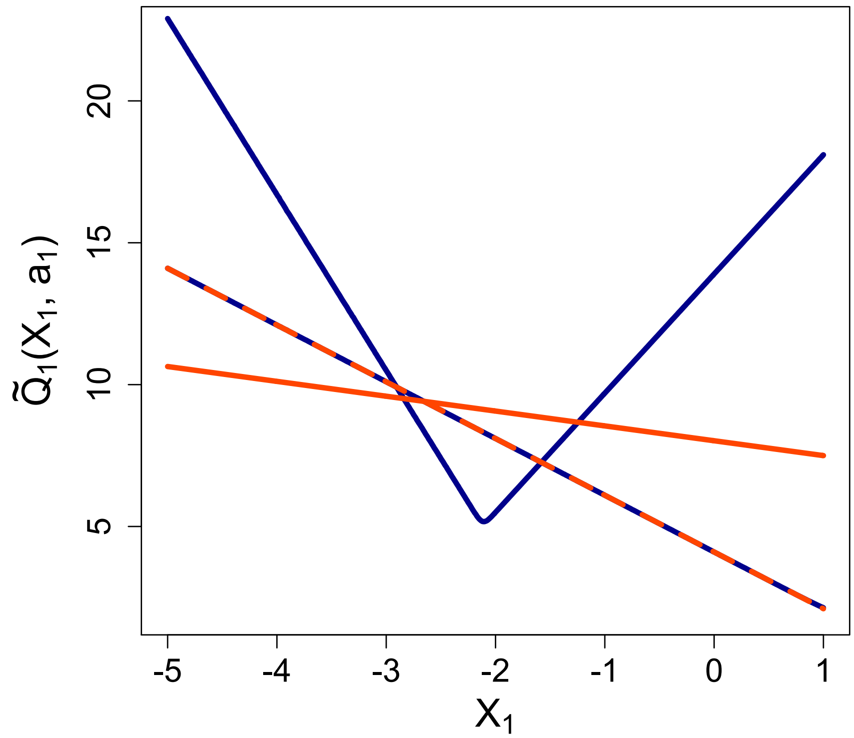

Fig. 23: Blue dashed and solid lines represent the true first-stage Q-function evaluated at $A_1 = 1$ and $A_1 = -1$, respectively, i.e., $\tilde{Q}_1(X_1, a_1) = Q_1(X_1, a_1)$. Orange dashed and solid lines represent estimated first-stage Q-function, i.e., $\tilde{Q}_1(X_1, a_1) = \hat{Q}_1(X_1, a_1)$, evaluated at $A_1 = 1$ and $A_1 = -1$, respectively, using Q-learning with linear models.

ure 21. The nonparametric IQ-learning estimator produced the lowest integrated mean squared error, however, this did not translate into any improvement in average value over the average value of Q-learning. The normal IQ-learning estimator displayed the poorest coverage in this case, but the nonparametric IQ-learning estimator came close to achieving the nominal level for all values of C.

6. Remark on Figure 3

The plot in the left frame of Figure 3 in Section 3 of the main paper gives a range of X_1 values and second-stage treatment effect sizes for which Q-learning with linear models does and does not agree with the true first-stage Q-function. Figure 23 is a plot of the true and Q-learning estimated first-stage Q-functions for the same range of X_1 values and for a single effect size, $C = 1$, where C is a constant that determines the effect size, defined in Section 3 of the main paper. The example in Figure 23 illustrates why the pattern of Figure 3 in the main paper is strange. Because higher values of the first-stage Q-function are desired, the true Q-function indicates patients presenting with X_1 below -3 and above -1.5 should be treated with $A_1 = -1$ and otherwise given $A_1 = 1$. However, the estimated first-stage Q-function using linear models cannot capture the non-linearity in $Q_1(X_1, a_1)$ and thus treats all patients presenting with X_1

below -3 with $A_1 = 1$, contrary to the true optimal treatment. In addition, the estimated Q-function treats patients presenting with X_1 between approximately -3 and -1.5 with $A_1 = 1$, contrary to the true optimal rule that treats these patients with $A_1 = -1$. Varying C results in different degrees of non-linearity in the true first-stage Q-function, resulting in the pattern observed in Figure 3 of Section 3 in the main paper.

7. WEB SUPPLEMENT F: APPLICATION TO STAR*D

Table 1: *Variables comprising patient trajectories in the STAR*D data analysis.*

Variable	Description
$X_{1,1} \in [0, 27]$	27 minus the baseline patient depression score.
$X_{1,2} \in \mathbb{R}$	Pre-randomization slope of patient depression score, computed by taking the difference between the measured depression score at study entry and the beginning of the first randomized stage. This difference is then divided by the time between study entry and first randomization. Negative values are associated with symptom improvement.
$A_1 \in \{-1, 1\}$	Initial treatment, coded so that $A_1 = 1$ corresponds to Selective Serotonin Reuptake Inhibitor and $A_1 = -1$ otherwise.
$Y_1 \in [0, 27]$	27 minus the patient depression score measured at the end of the first stage.
$R \in \{0, 1\}$	First-stage responder indicator. $R = 1$ indicates remission in stage one and exit from the study.
$X_{2,1} \in [0, 27]$	27 minus the patient depression score measured just prior to the second randomization.
$X_{2,2} \in \mathbb{R}$	First-stage slope of patient depression score, computed as the difference between the patient depression scores measured at the beginning and end of the first randomized stage. This difference is then divided by the time spent in the first randomized stage. Negative values are associated with symptom improvement.
$A_2 \in \{-1, 1\}$	Second stage treatment, coded so that $A_2 = 1$ corresponds to Selective Serotonin Reuptake Inhibitor and $A_2 = -1$ otherwise.
$Y_2 \in [0, 27]$	Second-stage outcome, defined as 27 minus the end of second-stage patient depression score.

Table 2: *Number of patients per treatment strategy by responder status.*

Treatment Sequence	Responders	Non-responders
(SSRI, NA)	319	NA
(non-SSRI, NA)	147	NA
(SSRI, SSRI)	NA	70
(SSRI, non-SSRI)	NA	120
(non-SSRI, SSRI)	NA	0
(non-SSRI, non-SSRI)	NA	139

References

Carroll, R. J. and Ruppert, D. (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer.

[*Received April* 2012. *Revised September* 2012]