Penalized Q-Learning for Dynamic Treatment Regimens

R. Song, W. Wang, D. Zeng and M. R. Kosorok

North Carolina State University, The University of Texas Health Science Center at Houston, and University of North Carolina

S1 Discussion of Simulation Studies

Setting 1 is a setting where there is no second-stage treatment effect, as $\psi_{20}^T S_{2(2)} = 0$ for all values of $S_{2(2)}$. The hard-max estimator will incur asymptotic biases for all the four terms $|f_1|$, $|f_2|$, $|f_3|$ and $|f_4|$, all four at about the same order of $\sqrt{2/(n\pi)} = 0.036$, as in this case $\{|f_i|\}_{i=1}^4 = 0$. As shown in (13), the biases in the estimation of $|f_1|$, $|f_2|$, $|f_3|$ and $|f_4|$ will be almost completely canceled out in the estimation of β_{12}^0 and ψ_{12}^0, due to the fact that the sum of the coefficients are zero. These biases of estimating $|f_1|$, $|f_2|$, $|f_3|$ and $|f_4|$ are largely canceled out in the estimation of ψ_{11}^0, as the sum of the coefficients are close to zero. The hard-max estimator of β_{11}^0 has a significant bias because the coefficients of the four absolute value terms, q_1, q_2, $0.5 - q_1$ and $0.5 - q_2$, are all positive and sum to 1.

The simulation results of Setting 1 are consistent with the theoretical observations in terms of the hard-max estimation. The oracle estimator automatically sets $\widehat{\psi}_2 = 0$. It has no significant bias, with standard errors accurately predicted by the theory and 95% confidence interval coverage close to the nominal value. The penalized Q-learning based estimator's performance is actually identical to the oracle estimator. The hard-max estimator has a significant bias and inferior mean square error in $\hat{\beta}_{11}$ while remaining consistent for estimation of the other three stage-1 parameters.

Setting 2 is regular but very close to Setting 1 with $\psi_{20}^T S_{2(2)}$ all equal to 0.01 for all values of $S_{2(2)}$. The hard-max estimator's performance is very similar to setting 1. Its 95% confidence interval shows poor coverage for β_{11}^0 and ψ_{11}^0. As the value of $\psi_{20}^T S_{2(2)}$ is nonzero, the oracle estimator reduces to the hard-max in this setting. Although the penalized Q-learning based estimator demonstrates a small bias (-0.009) in the estimation of β_{11}^0, the bias is less than one fifth of that of the oracle estimator and the mean square error is less than half of the oracle estimator. Its standard error estimate remains close to the empirical values.

Setting 3 is another setting where there is no second-stage treatment effect. The value of $\psi_{20}^T S_{2(2)}$ is equal to 0 when $A_1 = -1$ with probability one half. The hard-max estimator incurs bias on the order of $O(n^{-1/2})$ in the estimation of $|f_2|$ and $|f_4|$, but

not $|f_1|$ and $|f_3|$, as $f_2 = f_4 = 0$ and $f_1 = f_3 = 1$. As seen from (13), the hard-max estimation of β_{12}^0 and ψ_{12}^0 is still approximately unbiased, due to the canceling-out of the coefficients of the four absolute value terms. The estimation of β_{11}^0 is biased from the true value at approximately half of the bias of Setting 1, due to the values of the $|f_i|$'s and their coefficients. The estimation of ψ_{11}^0 is also biased, with similar magnitude of bias as in $\hat{\beta}_{11}$ but with reversed sign. The simulation study exactly confirms the theoretical observations of the hard-max estimator. The oracle estimator has no statistically significant bias and its standard error is precisely predicted by the theoretical calculation. The penalized Q-learning based estimator has a bias in $\hat{\psi}_{11}$ but the bias is still three times smaller than that of the hard-max estimator. Otherwise, the penalized Q-learning based estimator has almost exactly the same performance as the oracle estimator.

Setting 4 is a regular setting but very close to Setting 3. The hard-max estimator's performance is similar to Setting 3. The oracle estimator reduces to the hard-max estimator. The penalized Q-learning based estimator outperforms the oracle estimator, with both a smaller bias (5 times smaller), and a correctly predicted standard error. This phenomena is consistent with findings in Setting 2.

In Setting 5, the term $\psi_{20}^T S_{2(2)}$ is equal to zero when $(O_2, A_1) = (-1, -1)$ with probability one fourth. The hard-max estimator will incur bias in the estimation of $|f_4|$, since $f_4 = 0$. Consequently, all the four stage-1 parameter estimators will be biased. The bias in $\hat{\beta}_{11}$ will be approximately a quarter of that in Setting 1. The bias in $\hat{\beta}_{12}$ is about half of that of $\hat{\beta}_{11}$, with reversed sign. The bias in $\hat{\psi}_{11}$ is about the same magnitude as that of $\hat{\beta}_{11}$, with reversed sign. The bias in $\hat{\psi}_{12}$ is about half of that of $\hat{\beta}_{11}$. In this setting, the oracle estimator has the best performance, with no significant bias and well predicted standard errors. The penalized Q-learning based estimator has a bias in $\hat{\psi}_{11}$ but the bias is much smaller than the hard-max estimator. The penalized Q-learning based estimator has no noticeable bias in the other three parameter estimations and the standard error calculations are accurate when compared to Monte-Carlo errors.

Setting 6 is a completely regular setting with values of $\psi_{20}^T S_{2(2)}$ well above zero. The penalized Q-learning based estimator has almost identical performance as the oracle estimator, which is the same as the hard-max estimator. Both estimators are unbiased with accurately calculated standard errors.

In summary, the behavior of the PQ-estimator, including its bias, mean square error, theoretically computed standard error and coverage probability of theoretically computed 95% confidence intervals, are consistent in all six settings.

[1] proposed several bootstrapped confidence intervals for the hard-max estimator as well as hard-threshold estimators with α in Step 2' set to be 0.08 ($\text{HT}_{0.08}$) or 0.20 ($\text{HT}_{0.20}$) and the soft-threshold estimator (ST). In order to compare the proposed PQ-estimator confidence intervals with these bootstrapped methods, we re-ran the simulations with the PQ-estimator with sample size $n = 300$ and 1000 replications. The simulation results from different inferential methods in the six settings are compared in Table 3, where the results from the hard-max, hard threshold and soft threshold methods based on hybrid bootstrapping for variance estimation are shown. Overall, the other competing methods cannot provide consistent coverage rates across all six settings while our PQ-method

always gives coverage probabilities which are not significantly different from the nominal level.

S2 The Analysis of a Real Example

We here present the analysis of the mental health study data described in [2] using penalized Q-learning. The study is a prospective multi-site randomized clinical trial designed to determine the comparative effectiveness of different multi-level treatment options for patients with major depressive disorder. A total of 4041 patients with nonpsychotic major depressive disorder were enrolled and initially treated with citalopram for a minimum of 8 weeks, with strong encouragement to complete 12 weeks in order to maximize benefit. During this and all subsequent treatment levels, patients have clinic visits at weeks 0, 2, 4, 6, 9 and 12.

During all clinic visits, symptomatic status is measured by the 16-item Quick Inventory of Depressive Symptomatology Clinician-Rating scores. Patients who did not have a satisfactory response to treatment, defined as either $< 50\%$ reduction in the scores or the scores> 5, are eligible for Level 2 treatment. Seven treatment options are available at Level 2, which can be classified into two classes, (1) Medication or Psychotherapy Switch: sertraline, venlafaxine, bupropion or Cognitive Psychotherapy; and (2) Medication or Psychotherapy Augmentation: citalopram+bupropion, citalopram+buspirone or citalopram+Cognitive Psychotherapy. Patients who were assigned to Cognitive Psychotherapy or citalopram+Cognitive Psychotherapy in Level 2 and did not have a satisfactory response would be eligible for Level 2A, during which they would be treated with either venlafaxine or bupropion. Patients who did not respond satisfactorily at Level 2 and Level 2A, if applicable, would continue to Level 3 treatment. Level 3 includes four options: Medical Switch to mirtazapine or nortriptyline, and Medical augmentation with either lithium or thyroid hormone added to level 2 or 2A treatments. Patients who did not respond satisfactorily to Level 3 treatments would continue to Level 4 treatments, which include two options: switch to tranylcypromine or mirtazapine+venlafaxine. For a complete description of the study, see [2] and [4].

In this analysis, for demonstration purpose, we consider a subgroup of patients, the 112 patients who were randomized to either bupropion or sertraline in Level 2, had no satisfactory response at the end of Level 2, and were then randomized to either mirtazapine or nortriptyline in Level 3. The analysis focuses on selecting the optimized treatment regimen at Level 2 and Level 3, out of the 4 unique treatment combinations. Since the higher the score is, the more severe the symptom is, we define the clinical outcome as the negative of the score collected at the end of Level 3. Similarly as discussed in [3], the state variable used to tailor individual treatment is the changing rate of the score during the previous treatment level. We dichotomize the changing rates at zero. Two patients were further removed due to missing values in the clinical outcome or the tailoring variables. The tuning parameter $\lambda = 4$ under five-fold cross validation. The parameter ϕ in the adaptive lasso penalty takes value 2.

Table 1: Summary statistics and empirical coverage probability of 95% nominal percentile CIs for ψ_{11}^0 using the hard max (HM) estimator, the hard threshold estimator with $\alpha = 0.08$ ($HT_{0.08}$) and $\alpha = 0.20$ ($HT_{0.20}$), and the soft-threshold estimator quoting the simulation results from [1]. Specifically, "MSE" refers to the mean squares error, "Std-MC" refers to the standard deviation of 1000 estimates, "Std" refers to the average of the 1000 standard error estimates and "CP" refers to the empirical coverage probability of the 95% nominal percentile confidence interval. A "*" indicates a significantly different coverage rate from the nominal rate. "PB", "HB" and "DB" denote percentile bootstrap, hybrid bootstrap and double bootstrap, respectively.

Setting	Estimator	Bias	MSE	Std-MC	Std	CP(PB HB DB)		
1	HM	.0003	.0045	.066	–	96.8*	93.5*	93.6
	$HT_{.08}$	.0017	.0044	.066	–	97.0*	95.0	–
	$HT_{.20}$	.0002	.0050	.071	–	97.4*	92.8*	–
	ST	.0009	.0036	.060	–	95.3	96.1	–
	oracle	-.0015	.0034	.058	.058	94.6	–	–
	PQ	-.0013	.0036	.060	.061	95.1	–	–
2	HM	.0003	.0045	.065	–	96.7*	93.4*	93.6
	$HT_{.08}$	.0010	.0044	.066	–	97.1*	95.3	–
	$HT_{.20}$	.0003	.0050	.071	–	97.3*	93.5*	–
	ST	.0008	.0036	.060	–	95.4	95.9	–
	oracle	-.0025	.0043	.065	.075	97.5*	–	–
	PQ	-.0026	.0035	.059	.060	94.0	–	–
3	HM	-.0401	.0075	.075	–	88.4*	92.7*	94.8
	$HT_{.08}$	-.0083	.0059	.076	–	94.3	94.3	–
	$HT_{.20}$	-.0179	.0065	.079	–	93.5*	93.5*	–
	ST	-.0185	.0058	.074	–	93.4*	94.9	–
	oracle	-.0032	.0050	.071	.071	95.3	–	–
	PQ	-.0182	.0057	.073	.076	95.2	–	–
4	HM	-.0353	.0072	.076	–	89.6*	93.1*	94.4
	$HT_{.08}$	-.0037	.0058	.076	–	94.6	94.1	–
	$HT_{.20}$	-.0130	.0064	.079	–	93.9	92.8*	–
	ST	-.0138	.0057	.074	–	94.1	95.0	–
	oracle	-.0330	.0069	.076	.079	94.9	–	–
	PQ	-.0073	.0055	.074	.075	95.5	–	–
5	HM	-.0209	.0074	.077	–	92.7*	93.1*	94.2
	$HT_{.08}$	-.0059	.0071	.084	–	93.9	93.2*	–
	$HT_{.20}$	-.0101	.0073	.084	–	93.3*	93.0*	–
	ST	-.0065	.0069	.083	–	93.8	94.6	–
	oracle	-.0002	.0056	.075	.079	95.7	–	–
	PQ	-.0188	.0066	.079	.080	95.3	–	–
6	HM	.0009	.0067	.082	–	95.0	93.8	95.0
	$HT_{.08}$	.0003	.0081	.090	–	95.1	88.5*	–
	$HT_{.20}$	.0011	.0074	.086	–	94.8	91.2*	–
	ST	.0052	.0074	.086	–	94.8	91.7*	–
	oracle	.0003	.0061	.078	.080	95.4	–	–
	PQ	-.0012	.0062	.079	.080	95.3	–	–

Table 2: Level 3 regression model coefficient estimates using both unpenalized least squares estimation and individual penalized least squares estimation.

Variable	unpenalized		penalized	
	Coefficient	95% CI	Coefficient	95% CI
Intercept	-13.165	(-14.349, -11.981)	-13.185	(-14.330, -12.039)
O_1	-1.202	(-2.348, -0.057)	-1.124	(-2.233, -0.015)
A_1	0.004	(-0.945, 0.954)	-0.046	(-0.967, 0.874)
O_2	-0.587	(-1.605, 0.431)	-0.554	(-1.533, 0.425)
A_2	-1.276	(-2.460, -0.092)	-1.266	(-2.239, -0.292)
O_1A_2	-1.621	(-2.766, -0.475)	-1.300	(-2.410, -0.191)
A_1A_2	0.535	(-0.414, 1.484)	0.052	(-0.775, 0.880)
O_2A_2	0.278	(-0.740, 1.297)	0.017	(-0.748, 0.783)

Following the notations in the simulation study, let O_1 and O_2 be the indicator of whether the score changing rate is greater than zero in Level 1 and Level 2 respectively. Let $A_1 = 1$ if Level 2 treatment is sertraline and $A_1 = -1$ if it is bupropion. Let $A_2 = 1$ if Level 3 treatment is nortriptyline and $A_2 = -1$ if it is mirtazapine, and $R_2 =$ the negative score collected at the end of Level 3. The Level 3 regression model is:

$$R_2 = \beta_{21} + \beta_{22}O_1 + \beta_{23}A_1 + \beta_{24}O_2 + \psi_{21}A_2 + \psi_{22}A_2O_1 + \psi_{23}A_2A_1 + \psi_{24}A_2O_2 + \varepsilon_2.$$

Since the main effects of A_1 and O_2 are not statistically significant, we did not include additional interaction terms in the Level 3 model.

Table 2 shows the Level 3 regression model coefficient estimation using both unpenalized standard least squares estimation and individual penalized least squares estimation. Qualitatively, the unpenalized and penalized estimators are consistent. Patients whose symptoms worsened (i.e., $O_1 = 1$ or the score increased) during Level 1 would have a worse outcome. Level 2 treatments (sertraline versus bupropion) as well as the changing rate of the score during Level 2 show no differential effect on the final outcome. However, the two Level 3 treatment options show significantly different effects on patients with $O_1 = 1$ versus patients with $O_1 = -1$. Among patients whose symptoms worsened in Level 1, nortriptyline further worsened their symptom when compared to mirtazapine. Among patients whose symptom improved in Level 1, nortriptyline and mirtazapine show no obvious difference for the final outcome.

Quantitatively, the penalized estimator has smaller standard errors in the coefficient estimation of $\psi_2 = (\psi_{21}, \psi_{22}, \psi_{23}, \psi_{24})^T$ than the unpenalized estimator. In addition, the penalized estimator dramatically shrinks coefficients of the two unimportant predictors A_1A_2 and O_2A_2 toward zero. On the other hand, these two estimators are similar in the coefficient estimation of $\beta_2 = (\beta_{21}, \beta_{22}, \beta_{23}, \beta_{24})^T$, which is expected since the penalty is imposed only on ψ_2. In order to shrink coefficients of the unimportant predictors A_1 and O_2, one can further impose a penalty on $|\beta_2|$, which will not be implemented in this work. The lack of effect of A_1 and O_2 is actually expected since we include in this analysis only patients eligible for Level 3 treatment, in other words, only patients who did not respond satisfactorily to Level 2 treatment. This inclusion criteria is imposed because our current framework is built on the situation where all patients will be treated

Table 3: Values of $|\widehat{\psi}_2^T S_{2(2)}|$ in the data example.

			$|\widehat{\psi}_2^T S_{2(2)}|$	
O1	A1	O2	Unpenalized	Penalized
-1	-1	-1	0.468	0.035
-1	-1	1	0.088	0.000
-1	1	-1	0.601	0.070
-1	1	1	1.158	0.105
1	-1	1	3.153	2.601
1	1	-1	2.640	2.531
1	-1	-1	3.710	2.636
1	1	1	2.083	2.496

Table 4: Level 2 regression model coefficient estimation using both the Hard-max and the penalized Q-learning. PQ-learning: penalized Q-learning; CI: confidence interval.

	Hard-Max		PQ-learning	
Variable	Coefficient	Hybrid 95% CI	Coefficient	95% CI
Intercept	-11.063	(-12.482, -10.095)	-11.612	(-13.076, -10.149)
O_1	0.263	(-0.764, 1.547)	0.313	(-1.114, 1.740)
A_1	-0.119	(-1.120, 0.884)	-0.038	(-1.115, 1.039)
O_1A_1	-0.448	(-1.079, 0.251)	-0.085	(-0.830, 0.661)

in both stages. The extension to cases where patients may be cured during intermediate stages and hence not eligible for subsequent treatment stages is not trivial and will be considered in future work.

Table 3 shows the estimated values of $|\psi_2^T S_{2(2)}|$, where $S_{2(2)} = (1, O_1, A_1, O_2)^T$. When $O_1 = -1$, the Level 3 treatment effect is small but the unpenalized estimator shows significant bias from zero. On the other hand, the penalized estimator successfully shrinks the value of $|\psi_2^T S_{2(2)}|$ in all groups close to zero. Due to the limitation of the current local quadratic approximation algorithm, the penalized estimator cannot exactly set $|\psi_2^T S_{2(2)}|$ to zero, but the bias is significantly smaller than that of the unpenalized estimator. When $(O_1, A_1, O_2) = (-1, -1, 1)$, the penalized estimation of $|\psi_2^T S_{2(2)}|$ falls below the preselected cutoff of 0.001 and is shown as 0 in Table 3. When $O_1 = 1$, the treatment option mirtazapine can significantly improve the symptoms. Since A_1 and O_2 have no important effect on the outcome, we expect similar treatment effects among the four groups with $O_1 = 1$. From this point of view, the unpenalized estimator is inferior since it shows much bigger variation than the penalized estimator.

We next consider the Level 2 regression model. The pseudo-outcome $\widehat{Y}$ is defined as $\widehat{Y} = \beta_2^T S_{2(1)} + |\psi_2^T S_{2(2)}|$ and we impose the following Level 2 model:

$$\widehat{Y} = \beta_{11} + \beta_{12} O_1 + \beta_{13} A_1 + \beta_{14} O_1 A_1.$$

Table 4 shows the Level 2 model coefficient estimation using both the hard-max estimator and the penalized Q-learning based estimator. The coefficient estimation for the

intercept and O_1 are similar from two different estimation methods. In the estimation of coefficients for A_1 and O_1A_1, the penalized Q-learning based estimators are substantially closer to zero. Based on the 95% confidence intervals from penalized Q-learning, O_1 and A_1 have no effect on the pseudo-outcome. Since A_1 shows no effect in Level 3 regression either, it is easy to interpret its lack of effect on the pseudo-outcome. In contrast, O_1 is a strong predictor in Level 3 treatment. Its lack of effect in Level 2 regression may be explained as follows. In Level 3 regression, the $O_1 = 1$ group's clinical outcome is smaller than the $O_1 = -1$ group's clinical outcome by $2\beta_{22} \approx 2.24$. However, the optimal Level 3 treatment can increase the $O_1 = 1$ group's clinical outcome by $|\psi_2^T S_{2(2)}| \approx 2.6$ but cannot increase the $O_1 = -1$'s clinical outcome. Hence O_1 has no net effect on the pseudo-outcome.

Our analysis found that the optimal Level 2 and Level 3 treatment regimen in this subgroup of patients is the following. In Level 2, there is no difference in choosing sertraline or bupropion. If a patient's symptom worsens in Level 1 and remains unsatisfactory in Level 2, mirtazapine is a better option for Level 3 treatment when compared to nortriptyline. If a patient's symptom improves in Level 1 and remains unsatisfactory in Level 2, mirtazapine or nortriptyline have a similar effect as a Level 3 treatment.

S3 Proofs

Proof of Theorem 1.

Let $\alpha_n = C(1/\sqrt{n} + a_n)$, where C is a constant to be determined later. We aim to show that for any given $\epsilon > 0$, there exists a large constant C such that

$$P(\inf_{\|u\|=1} W_2(\theta_{20} + \alpha_n u) > W_2(\theta_{20})) \geq 1 - \epsilon, \tag{S3.1}$$

where $u = (u_1^T, u_2^T)^T$, $u_1 \in \mathbb{R}^p$, $u_1 \in \mathbb{R}^q$ with $\|u\| = 1$, and

$$W_2(\theta_2) = \sum_{i=1}^{n} (R_{2i} - Q_2(S_{2i}, A_{2i}; \theta_2))^2 + \sum_{i=1}^{n} p_{\lambda_n}(|\psi_2^T S_{2i(2)}|).$$

Let $G_{n2}(\theta_2) = \sum_{i=1}^{n} (R_{2i} - Q_2(S_{2i}, A_{2i}; \theta_2))^2$. Using $p_{\lambda_n}(0) = 0$, we have

$$\begin{aligned} D_n(u) &= W_2(\theta_{20} + \alpha_n u) - W_2(\theta_{20}) \\ &= G_{n2}(\theta_{20} + \alpha_n u) - G_{n2}(\theta_{20}) + \sum_{i=1}^{n} p_{\lambda_n}(|(\psi_{20} + \alpha_n u_2)^T S_{2i(2)}|) - \sum_{i=1}^{n} p_{\lambda_n}(|\psi_{20}^T S_{2i(2)}|) \\ &\geq G_{n2}(\theta_{20} + \alpha_n u) - G_{n2}(\theta_{20}) + \sum_{k=1}^{K_1} n_k p_{\lambda_n}(|(\psi_{20} + \alpha_n u_2)^T v_k|) - \sum_{k=1}^{K_1} n_k p_{\lambda_n}(|\psi_{20}^T v_k|). \end{aligned}$$

By Taylor expansion of W_2 and noting that $\nabla^2_{\theta_2\theta_2} G_{n2}(\theta_{20}) \geq 1/2 I_{20}$ by condition

B3 for large n, we obtain

$$\begin{aligned} D_n(u) \geq \quad & \alpha_n G'_{n2}(\theta_{20})u + n/2 I_{20}\alpha_n^2\{1+o_p(1)\} \\ & -c_0 n\sum_{k=1}^{K_1}\frac{n_k}{n}\alpha_n u_2^T p'_{\lambda_n}(|\psi_{20}^T v_k|) - c_0^2 n\sum_{k=1}^{K_1}\frac{n_k}{n}\alpha_n^2 u_2^T p''_{\lambda_n}(|\psi_{20}^T v_k|)u_2\{1+o(1)\}, \end{aligned}$$

where c_0 is a finite upper bound for $\|S_{2(2)}\|$.

Since θ_{20} minimizes the limit of $G_{n2}(\theta_2)$, it is easy to see that

$$G'_{n2}(\theta_{20}) = -2n(\mathbb{P}_n - \mathbb{P})\left[\nabla_{\theta_2} Q_2(S_2, A_2; \theta_{20})(Y_2 - Q_2(S_2, A_2; \theta_{20}))\right] = O_p(n^{1/2}),$$

where $\mathbb{P}f$ is $\lim_n \mathbb{P}_n f$ for a function f and $\mathbb{P}_n f = 1/n\sum_{i=1}^n f(X_i)$ is the empirical function for independent identically distributed random variable X_i, $i = 1, \ldots, n$. Thus,

$$-\alpha_n G'_{n2}(\theta_{20})u \leq n/4 I_{20}\alpha_n^2 + O(1).$$

Moreover, according to the property of the penalty function,

$$\sum_{k=1}^{K_1}\frac{n_k}{n}\alpha_n u_2^T p'_{\lambda_n}(|\psi_{20}^T v_k|) \leq O(1)a_n\alpha_n,$$

and

$$\sum_{k=1}^{K_1}\frac{n_k}{n}\alpha_n^2 u_2^T p''_{\lambda_n}(|\psi_{20}^T v_k|)u_2 = o(\alpha_n^2).$$

We conclude that

$$D_n(u) \geq n(1/8 I_{20}\alpha_n^2 - O(n^{-1}) - O(a_n)\alpha_n).$$

Therefore, if we choose the constant C large enough in $\alpha_n = C(1/\sqrt{n} + a_n)$, the right-hand side of the above inequality is strictly positive, which implies that there exists a local minimizer $\widehat{\theta}_2$ such that $\|\widehat{\theta}_2 - \theta_{20}\| = O(n^{-1/2} + a_n)$. This concludes the proof of Theorem 1. □

Proof of Theorem 2.

We consider the sets in the probability space:

$$\mathcal{C}_k = \{\psi_{20}^T v_k = 0, \widehat{\psi}_2^T v_k \neq 0\}, \quad k = K_1 + 1, ..., K.$$

We will show that for any $\epsilon > 0$, when n is large enough, $P(\mathcal{C}_k) < \epsilon$.

Since $\widehat{\psi}_2^T v_k = O_p(n^{-1/2})$ by Theorem 1, for any $\epsilon > 0$, there exists some M such that for sufficiently large n,

$$P(\mathcal{C}_k) < \epsilon/2 + P(\psi_{20}^T v_k = 0, \widehat{\psi}_2^T v_k \neq 0, |\widehat{\psi}_2^T v_k| < M n^{-1/2}).$$

After differentiating $W_{n2}(\theta_2)$ with respect to ψ_2, we obtain

$$n^{-1/2}\nabla_{\psi_2}G_{n2}(\widehat{\theta}_2) + n^{-1/2}\sum_{i=1}^{n} p'_{\lambda_n}(|\widehat{\psi}_2^T S_{2i(2)}|)\text{sgn}(\widehat{\psi}_2^T S_{2i(2)})S_{2i(2)} = 0, \tag{S3.2}$$

where $\nabla_{\psi_2}G_{n2}$ is the score equation of $G_{n2}(\theta_2)$ with respect to ψ_2. Multiplying $\widehat{\psi}_2^T$ on both sides of (S3.2) yields that

$$n^{-1/2}\widehat{\psi}_2^T\nabla_{\psi_2}G_{n2}(\widehat{\theta}_2) + n^{-1/2}\sum_{i=1}^{n} p'_{\lambda_n}(|\widehat{\psi}_2^T S_{2i(2)}|)|\widehat{\psi}_2^T S_{2i(2)}| = 0.$$

Since $S_{2i(2)}$ takes values $v_1, ..., v_K$, the above equation can be rewritten as

$$n^{-1/2}\widehat{\psi}_2^T\nabla_{\psi_2}G_{n2}(\widehat{\theta}_2) + \sqrt{n}\sum_{k=1}^{K}\left[\frac{n_k}{n}p'_{\lambda_n}(|\widehat{\psi}_2^T v_k|)|\widehat{\psi}_2^T v_k|\right] = 0, \tag{S3.3}$$

where $n_k = \sum_{i=1}^{n} I(S_{2i(2)} = v_k)$. From the consistency of $\widehat{\psi}_2$, it is easy to verify that the first term in the left-hand side of (S3.3) is $O_p(1)$. Moreover, for $k = 1, ..., K_1$, since $\widehat{\psi}_2^T v_k$ converges to $\psi_{20}^T v_k$, which is bounded away from zero, we conclude from property A1 that

$$\frac{n_k}{n}p'_{\lambda_n}(|\widehat{\psi}_2^T v_k|)|\widehat{\psi}_2^T v_k| \to 0.$$

Therefore, from (S3.3), there exists a constant m such that $P(\mathcal{D}_m) > 1 - \epsilon/4$, where the set $\mathcal{D}_m$ is defined as

$$\mathcal{D}_m = \left\{\sqrt{n}\sum_{k=K_1+1}^{K}\frac{n_k}{n}p'_{\lambda_n}(|\widehat{\psi}_2^T v_k|)|\widehat{\psi}_2^T v_k| < m\right\}.$$

Consequently,

$$P(\mathcal{C}_k) \le \epsilon/2 + P(\mathcal{D}_m^c) + P(\mathcal{D}_m \cap \left\{\widehat{\psi}_2^T v_k \ne 0, \widehat{\psi}_2^T v_k < Mn^{-1/2}\right\})$$

$$\le 3\epsilon/4 + P(\sqrt{n}p'_{\lambda_n}(|\widehat{\psi}_2^T v_k|)|\widehat{\psi}_2^T v_k| < 2m/p_k, \widehat{\psi}_2^T v_k \ne 0, \widehat{\psi}_2^T v_k < Mn^{-1/2}).$$

However, from property A2 of the penalty function, the second probability on the right-hand side will eventually be zero for n large enough. We thus conclude that when n is large enough, $P(\mathcal{C}_k) < \epsilon$. This proves Theorem 2. □

Proof of Theorem 3.

We perform Taylor expansion for the left-hand side of (S3.2) and also for the equation for $\widehat{\beta}_2$: $\nabla_{\beta_2}G_{n2}(\widehat{\theta}_2) = 0$. Note that from Theorem 2, with probability tending to one, $\text{sgn}(\widehat{\psi}_2^T S_{2i(2)}) = \text{sgn}(\psi_{20}^T S_{2i(2)})$. Thus,

$$p'_{\lambda_n}(|\widehat{\psi}_2^T S_{2i(2)}|)\text{sgn}(\widehat{\psi}_2^T S_{2i(2)})S_{2i(2)} = p'_{\lambda_n}(|\psi_{20}^T S_{2i(2)}|)\text{sgn}(\psi_{20}^T S_{2i(2)})S_{2i(2)}$$

$$+(p''_{\lambda_n}(|\psi_{20}^T S_{2i(2)}|)S_{2i(2)}S_{2i(2)}^T + o_p(1))(\widehat{\psi}_2 - \psi_{20}).$$

Hence, it holds that

$$\begin{aligned} 0 &= \nabla_{\theta_2} G_{n2}(\theta_{20}) + \nabla^2_{\theta_2\theta_2} G_{n2}(\theta_{20})^T(\widehat{\theta}_2 - \theta_{20}) + \sum_{i=1}^{n} p'_{\lambda_n}(|\psi_{20}^T S_{2i(2)}|)\mathrm{sgn}(\psi_{20}^T S_{2i(2)})S_{2i(2)} \\ &\quad + \Big(\sum_{i=1}^{n} p''_{\lambda_n}(|\psi_{20}^T S_{2i(2)}|)S_{2i(2)}S_{2i(2)}^T + o_p(1)\Big)(\widehat{\psi}_2 - \psi_{20}). \end{aligned}$$

Theorem 3 then follows from Slutsky's theorem and the central limit theorem. It also yields that $\widehat{\theta}_2$ is an asymptotically linear estimator for θ_{20} with influence function

$$F_2(\theta_{20}) = (F_{21}(\beta_{20})^T, F_{22}(\psi_{20})^T)^T = (I_{20}+\Sigma)^{-1}\nabla_{\theta_2} Q_2(S_2, A_2; \theta_{20})(R_2 - Q_2(S_2, A_2; \theta_{20})). \square$$

Proof of Theorem 4.

Utilizing the same expansion used in proving Theorem 3, we obtain

$$\sqrt{n} I_{10}(\widehat{\theta}_1 - \theta_{10}) = \sqrt{n}(\mathbb{P}_n - \mathbb{P})\left[\nabla_{\theta_1} Q_1(S_1, A_1; \theta_{10})(\widehat{Y}_1 - Q_1(S_1, A_1; \theta_{10}))\right]$$

$$-\sqrt{n}\mathbb{P}_n\left[\nabla_{\theta_1} Q_1(S_1, A_1; \theta_{10})(\widehat{Y}_1 - Y_1)\right] + o_p(1),$$

where $\widehat{Y}_1 = R_1 + \widehat{\beta}_{20}^T S_{2(1)} + |\widehat{\psi}_2^T S_{2(2)}|$ and $Y_1 = R_1 + \beta_{20}^T S_{2(1)} + |\psi_{20}^T S_{2(2)}|$.

On the other hand, with probability tending to one, $\widehat{\psi}_2^T S_{2(2)}$ has the same sign as $\psi_{20}^T S_{2(2)}$ from Theorem 2. Thus,

$$|\widehat{\psi}_2^T S_{2(2)}| - |\psi_{20}^T S_{2(2)}| = \mathrm{sgn}(\psi_{20}^T S_{2(2)})(\widehat{\psi}_2 - \psi_{20})^T S_{2(2)}.$$

Combining these results and using the fact that $\left\{|\psi^T S_{2(2)}|\right\}$ is a Donsker class, we obtain

$$\begin{aligned} &\sqrt{n}(\widehat{\theta}_1 - \theta_{10}) \\ = \;& \sqrt{n} I_{10}^{-1}(\mathbb{P}_n - \mathbb{P})\left\{\nabla_{\theta_1} Q_1(S_1, A_1; \theta_{10})(R_1 - Q_1(S_1, A_1; \theta_{10}))\right\} \\ & -\sqrt{n} I_{10}^{-1}\mathbb{P}\left[\nabla_{\theta_1} Q_1(S_1, A_1; \theta_{10})S_{2(1)}^T\right](\mathbb{P}_n - \mathbb{P})F_{21}(\beta_{20}) \\ & -\sqrt{n} I_{10}^{-1}\mathbb{P}\left[\nabla_{\theta_1} Q_1(S_1, A_1; \theta_{10})\mathrm{sgn}(\psi_{20}^T S_{2(2)})S_{2(2)}^T\right](\mathbb{P}_n - \mathbb{P})F_{22}(\psi_{20}) + o_p(1), \end{aligned}$$

where $F_{21}(\beta_{20})$ and $F_{22}(\psi_{20})$ are the respective influence functions for $\widehat{\beta}_2$ and $\widehat{\psi}_2$ as given in Theorem 3. The asymptotic normality of $\widehat{\theta}_1$ thus follows. $\square$

References

[1] Chakraborty, B., Murphy, S. & Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research* **19**, 317–343.

[2] Fava, M., Rush, A., Trivedi, M., Nierenberg, A., Thase, M., Sackeim, H., Quitkin, F., Wisniewski, S., Lavori, P., Rosenbaum, J., Kupfer, D. & STAR D Invest Grp (2003). Background and rationale for the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study. *Psychiatric Clinics of North America* **26**, 457–494.

[3] Pineau, J., Bellernare, M. G., Rush, A. J., Ghizaru, A. & Murphy, S. A. (2007). Constructing evidence-based treatment strategies using methods from computer science. *Drug and Alcohol Dependence* **88**, S52–S60.

[4] Rush, A., Fava, M., Wisniewski, S., Lavori, P., Trivedi, M., Sackeim, H., Thase, M., Nierenberg, A., Quitkin, F., Kashner, T., Kupfer, D., Rosenbaum, J., Alpert, J., Stewart, J., McGrath, P., Biggs, M., Shores-Wilson, K., Lebowitz, B., Ritz, L., Niederehe, G. & STAR D Investigators Grp (2004). Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Controlled Clinical Trials* **25**, 119–142.