Electronic Journal of Statistics
Vol. 8 (2014) 1290–1300
ISSN: 1935-7524
DOI: 10.1214/14-EJS905

Comment on "Dynamic treatment regimes: Technical challenges and applications"*†

Yair Goldberg

Department of Statistics, University of Haifa
Mount Carmel, Haifa 31905, Israel
e-mail: ygoldberg@stat.haifa.ac.il

Rui Song

Department of Statistics, North Carolina State University
Raleigh, NC 27695, USA
e-mail: rsong@ncsu.edu

Donglin Zeng and Michael R. Kosorok

Department of Biostatistics
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA
e-mail: dzeng@email.unc.edu; kosorok@unc.edu

Abstract: Inference for parameters associated with optimal dynamic treatment regimes is challenging as these estimators are nonregular when there are non-responders to treatments. In this discussion, we comment on three aspects of alleviating this nonregularity. We first discuss an alternative approach for smoothing the quality functions. We then discuss some further details on our existing work to identify non-responders through penalization. Third, we propose a clinically meaningful value assessment whose estimator does not suffer from nonregularity.

Received May 2014.

1. Introduction

The authors are to be congratulated for their excellent and thoughtful paper on statistical inference for dynamic treatment regimens. They have addressed several important and long-standing issues in this area. As discussed by the authors, nonsmoothness of the problem in some of the parameters of interest leads to estimators that are not smooth in the data. This in turn makes inference for these parameters challenging. In the following, we comment on a few additional strategies to alleviate the resulting nonregularity due to nonsmoothness.

*The first author was funded in part by ISF grant 1308/12. The other authors were funded in part by grant P01 CA142538 from the National Cancer Institute. The second author was also funded in part by NSF-DMS 1309465.

†Main article 10.1214/14-EJS920.

First, we discuss replacing the nonsmooth objective functions via a SoftMax Q-learning approach, which directly addresses the trade-off between bias and variance of the maximum operation in the local asymptotic framework. Proofs are given in the Appendix.

Nonregularity of the estimators for the parameters associated with the optimal treatment regimes is mainly due to the existence of non-responders to treatments. Therefore, it would be useful and important if we could identify these non-responders. In the second part, we review our existing work on non-responder identification via penalization. We also discuss how this penalization can alleviate, although not solve, some regularity issues.

For the third and final aspect we wish to discuss, we note that in some public health settings, the parameters in the dynamic treatment regime are not as important as the value function which reflects the overall population impact of the estimated regime and is perhaps the most important quantity to focus on for public health policy. We propose a truncated value function which only focuses on those subjects who are expected to have large treatment effects. We claim that this alternative value function is clinically meaningful and does not suffer from nonregularity.

2. SoftMax Q-learning

In this section we study the effect of replacing the max operator with a smoother version of it in the two-stage Q-learning algorithm discussed by Laber et al. We show that this smoothing can reduce the bias and can be controlled under local alternatives. The proposed SoftMax approach also sheds light on the bias/variance tradeoff which can be obtained by using over/under smoothing. In what follows, we briefly describe the SoftMax Q-learning algorithm, and then present some theoretical and simulation results.

2.1. Proposed algorithm

Consider the Q-learning algorithm discussed by Laber et al. in Section 2. In step 2 of the algorithm, the stage outcome is predicted by

$$\widetilde{Y} = \max_{a_2} Q_2(H_2, a_2; \hat{\beta}_2).$$

We propose replacing $\widetilde{Y}$ with a SoftMax version of it. Define the SoftMax function by (see Fig. 1)

$$SoftMax(x, y, \alpha) = \frac{1}{\alpha} \log \{e^{\alpha x} + e^{\alpha y}\}, \quad \alpha > 0.$$

Let

$$\breve{Y} = SoftMax\left(Q_2(H_2, a_{2,1}; \hat{\beta}_2), Q_2(H_2, a_{2,2}; \hat{\beta}_2), \alpha\right)$$

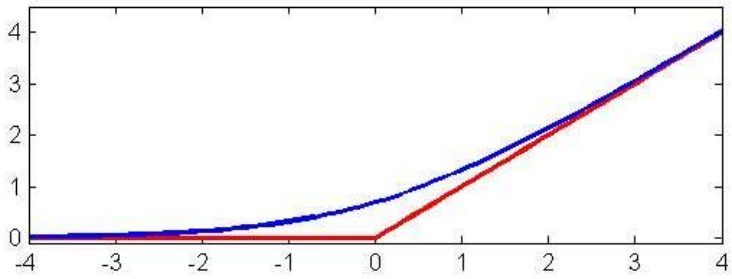

FIG 1. *The function* $\log\{\exp(x)+\exp(0)\}$ *in blue and the function* $\max(x,0)$ *in red. Note that the functions roughly agree for* $x \notin [-3,3]$.

$$= \frac{1}{\alpha}\log\left\{e^{\alpha H'_{2,0}\hat{\beta}_{2,0}} + e^{\alpha\left(H'_{2,0}\hat{\beta}_{2,0}+H'_{2,1}\hat{\beta}_{2,1}\right)}\right\}$$
$$= H'_{2,0}\hat{\beta}_{2,0} + \frac{1}{\alpha}\log\left\{1+e^{\alpha H'_{2,1}\hat{\beta}_{2,1}}\right\}.$$

The estimator $\hat{\beta}_1$ of β_1 is given by $\hat{\Sigma}_1^{-1}\mathbb{P}_n B_1 \breve{Y}$. We note that the algorithm discussed by Laber et al. is obtained as the limit, as α goes to infinity, of the SoftMax Q-learning algorithm discussed here.

2.2. *Theory*

In the following we briefly discuss the asymptotic properties of $\hat{\beta}_1$. We first discuss the limiting distribution of $\sqrt{n}(\hat{\beta}_1 - \beta_1^*)$. We then discuss this limiting distribution under local alternatives. Finally, we discuss the asymptotic bias. The proofs appear in the Appendix.

Theorem 1. *Assume (A1)–(A2) from Laber et al., and let* $\alpha_n \to \infty$ *such that* $\sqrt{n}/\alpha_n \to a_\infty$ *for* $a_\infty \in [0,\infty)$. *Then*
(i) If $a_\infty = 0$,

$$\sqrt{n}(\hat{\beta}_1 - \beta_1^*) \rightsquigarrow \mathbb{S}_\infty + \Sigma_{1,\infty}^{-1} P\left(\mathbb{T}_\infty\right).$$

(ii) If $0 < a_\infty < \infty$, *then*

$$\sqrt{n}(\hat{\beta}_1 - \beta_1^*) \rightsquigarrow \mathbb{S}_\infty + \Sigma_{1,\infty}^{-1} P\left(\mathbb{T}_\infty\right) + a_\infty \log(2)\Sigma_{1,\infty}^{-1} P B_1 \mathbf{1}\left\{H'_{2,1}\beta_{2,1}^* = 0\right\},$$

where

$$\mathbb{T}_\infty = B_1\left(H'_{2,1}\mathbb{V}_\infty \mathbf{1}\left\{H'_{2,1}\beta_{2,1}^* > 0\right\} + \frac{1}{2}H'_{2,1}\mathbb{V}_\infty \mathbf{1}\left\{H'_{2,1}\beta_{2,1}^* = 0\right\}\right).$$

For local alternatives the limiting distribution is given below.

Theorem 2. *Assume (A1)–(A3) from Laber et al., and let* $\alpha_n \to \infty$ *such that* $\sqrt{n}/\alpha_n \to a_\infty$ *for* $a_\infty \in (0,\infty)$. *Then*

$$\sqrt{n}(\hat{\beta}_1 - \beta_1^*) \rightsquigarrow \mathbb{S}_\infty + \Sigma_{1,\infty}^{-1} P\left(\mathbb{T}_\infty\right) + \Sigma_{1,\infty}^{-1} P\left(\mathbb{W}_\infty\right),$$

where

$$\mathbb{T}_\infty = B_1 \left(H_{2,1}' \mathbb{V}_\infty \mathbf{1}\{H_{2,1}' \beta_{2,1}^* > 0\} + H_{2,1}' \mathbb{V}_\infty \left[a_\infty^{-1} H_{2,1}' s \right]_+ \mathbf{1}\{H_{2,1}' \beta_{2,1}^* = 0\} \right)$$
$$\mathbb{W}_\infty = B_1 \left(a_\infty \log \left\{ 1 + e^{a_\infty^{-1} H_{2,1}' s} \right\} - \left[H_{2,1}' s \right]_+ \right) \mathbf{1}\{H_{2,1}' \beta_{2,1}^* = 0\}.$$

The bound of the bias, scaled by root-n, under both standard and local alternatives asymptotics, is given below.

Corollary 1. *Let* $\text{Bias}(\hat{\beta}_1, c)$ *and* $\text{Bias}(\hat{\beta}_1, c, s)$ *be defined as in Laber et al.. Assume (A1)–(A2) from Laber et al., and let* $\alpha_n \to \infty$ *such that* $\sqrt{n}/\alpha_n \to a_\infty$ *for* $a_\infty \in (0, \infty)$. *Fix* $c \in \mathbb{R}^{p_{21}}$. *Then*

$$\text{Bias}(\hat{\beta}_1, c) \leq a_\infty \|\Sigma_{1,\infty}^{-1}\| P \| B \| \mathbf{1}\{|H_{2,1}' \beta_{2,1}^*| = 0\} + o_p(1).$$

When (A3) from Laber et al. also holds, then

$$\sup_{s \in \mathbb{R}^{p_{21}}} \text{Bias}(\hat{\beta}_1, c, s) \leq a_\infty \|\Sigma_{1,\infty}^{-1}\| P \| B \| \mathbf{1}\{|H_{2,1}' \beta_{2,1}^*| = 0\} + o_p(1).$$

The above results show that by choosing the scale of α, the bias can be controlled. Theorem 2 shows that this control of the bias directly influences the variance, at least under local alternatives.

For inference, we need to discuss two different settings. When holding α fixed, as n goes to infinity, standard inference for the parameters is valid, as the problem becomes regular. However, this comes with the price that the bias does not vanish even asymptotically (see also the discussion in Section 4). As proved in Theorem 2, when taking α to infinity, as n goes to infinity, the problem is nonregular. Thus, adaptive confidence intervals, such as the one suggested by Laber et al., are needed in order to perform valid inference.

2.3. Simulations for SoftMax

We compare the small-sample behaviour of SoftMax to that of soft-threshholding using the example setting discussed in Section 3 of Laber et al. Let $\theta^* = \max(\mu_0^*, \mu_1^*)$. The max estimator is defined by

$$\hat{\theta} \equiv \max(\hat{\mu}_0, \hat{\mu}_1) = \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} + \frac{|\hat{\mu}_0 - \hat{\mu}_1|}{2}.$$

A soft-thresholdling estimator is defined by

$$\hat{\theta}^\sigma = \frac{\hat{\mu}_0 + \hat{\mu}_1}{2} + \frac{|\hat{\mu}_0 - \hat{\mu}_1|}{2} \left(1 - \frac{4\sigma}{n(\hat{\mu}_0 - \hat{\mu}_1)} \right).$$

Finally, the SoftMax estimator is defined by

$$\breve{\theta} = \mathit{SoftMax}(\hat{\mu}_0, \hat{\mu}_1, \alpha).$$

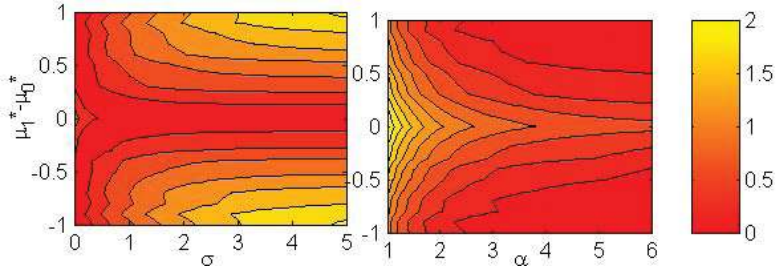

FIG 2. ***Left:*** *Bias for soft-thresholding.* ***Right:*** *Bias for SoftMax. In both panels the bias is measured in units of* $1/\sqrt{n}$ *for* $n = 10$*, as a function of effect size and of the tuning parameters,* σ *and* α*, for the soft-thresholding and SoftMax, respectively.*

Let $Y|A \sim N(\mu_a, 1)$, $a = 0, 1$, and assume that the treatment assignment is perfectly balanced. We use 1000 Monte Carlo replicates to estimate the bias for each parameter setting. Figure 2 below shows the bias as a function of the treatment effect $\mu_1^* - \mu_0^*$ and with tuning parameters $\sigma \in [0, 5]$ and $\alpha \in [1, 6]$ for the soft-thresholding and SoftMax, respectively. It appears that the SoftMax does not suffer from large bias on points away from $\mu_1^* - \mu_0^* = 0$. Also, as expected from Theorem 1, the bias decreases as α increases.

3. Penalized and adaptive Q-learning

In Penalized Q-learning (Song et al., 2011) and adaptive Q-learning (Goldberg et al., 2013), penalties were imposed on the term $H'_{2,1}\beta_{2,1}$ for each individual. This use of penalized estimation allows us to simultaneously estimate the second stage parameters and select individuals whose value functions are not affected by treatments, i.e., those individuals whose true values of $H'_{2,1}\beta^*_{2,1}$ are zero. Although the penalized method does not solve the non-regularity issue in estimating β's, our numerical studies have demonstrated that penalized Q-learning is not only able to reduce bias, but also provides better coverage of confidence intervals in a number of scenarios, as compared to the hard thresholding method of Moodie and Richardson (2010) and some soft thresholding methods including resampling approaches. Furthermore, the inference approach for penalized methods described in Zhang and Zhang (2014) appears to be able to handle diverging model perturbations. Finally, a nice feature of our penalized learning is that it enables us to identify non-responders, who may also have small treatment benefits even under a local alternative. Since it is clinically and practically most useful to target groups whose treatment benefit is large, identifying those subjects with small treatment benefits is useful for better allocation of resources and for reducing costs.

4. Truncated value function

The non-regularity issue arises primarily in settings where there are some subjects who do not respond to treatments at the second stage and where inference focuses on effect size. In the context of public health policy, we think that (i) the overall benefit (value) may be of greater interest compared to individual effect sizes and (ii) those subjects who are not sensitive to treatments (approximate non-responders) should not have a large impact on the overall decision making process. Thus, we propose an appropriate alternative criterion, namely the ϵ-truncated value, for evaluating the optimal policy as follows:

$$V_\epsilon(d_1, d_2) = E_d\left[(Y_1(d_1) + Y_2(d_2))I(\delta(X_1) > \epsilon, \delta(X_2) > \epsilon)\right],$$

where $\delta(X_1)$ and $\delta(X_2)$ denote the expected treatment effects at the first and second stages respectively. Here, ϵ is a small constant indicating a clinically meaningful effect size.

Under a SMART trial with randomization probabilities π_k at stage k ($k = 1, 2$), this truncated value is equal to

$$E\left[(Y_1 + Y_2)I(A_1 = d_1(X_1), A_2 = d_2(X_2), \delta(X_1) > \epsilon, \delta(X_2) > \epsilon)/(\pi_1\pi_2)\right].$$

Compared to the usual value function, we can see that $V_\epsilon(d_1, d_2)$ differs by at most $O(\epsilon)$. Using the Q-learning model, the above value function for the estimated rule is

$$V_\epsilon(\hat{d}_1, \hat{d}_2) = E\left[(Y_1 + Y_2)I(A_1\hat{\beta}_1X_1 > 0, A_2\hat{\beta}_2X_2 > 0, |\hat{\beta}_1X_1| > \epsilon, |\hat{\beta}_2X_2| > \epsilon)/(\pi_1\pi_2)\right].$$

One advantage of considering this value function is that non-regularity will no longer be an issue since we have excluded the non-responders from the above statistic. One can easily show $\sqrt{n}(\hat{V}_\epsilon(\hat{d}_1, \hat{d}_2) - V_\epsilon(d_1, d_2))$ converges to the same normal distribution under local alternatives whether $P(\beta_2X_2 = 0) > 0$ or not.

5. Concluding remarks

We again thank the authors for their very interesting work which likely stimulate additional future research on this crucial topic. It is clear that there are many fundamental and unresolved computational, methodological and theoretical challenges remaining which will benefit from many diverse problem solving approaches. We look forward to seeing this intriguing research area continue to develop.

Appendix: Proofs

Sketch of proof of Theorem 1. Using the same arguments that lead to Eq. 2 in Laber et al., we have

$$\sqrt{n}(\hat{\beta}_1 - \beta_1^*) = \hat{\Sigma}_1^{-1}\mathbb{P}_nB_1\left(\check{Y} - B_1'\beta_1^*\right) = \mathbb{S}_n + \hat{\Sigma}_1^{-1}\mathbb{P}_nB_1\mathbb{U}_n,$$

where $\mathbb{S}_n$ is smooth and asymptotically normal and

$$\mathbb{U}_n = \sqrt{n}\left(\frac{1}{\alpha_n}\log\left\{1+e^{\alpha_n H_{2,1}'\hat{\beta}_{2,1}}\right\} - \left[H_{2,1}'\beta_{2,1}^*\right]_+\right).$$

Note that

$$\begin{aligned}\mathbb{U}_n &= \sqrt{n}\left(\frac{1}{\alpha_n}\log\left\{1+e^{\alpha_n H_{2,1}'\hat{\beta}_{2,1}}\right\} - \frac{1}{\alpha_n}\log\left\{1+e^{\alpha_n H_{2,1}'\beta_{2,1}^*}\right\}\right)\\ &\quad + \sqrt{n}\left(\frac{1}{\alpha_n}\log\left\{1+e^{\alpha_n H_{2,1}'\beta_{2,1}^*}\right\} - \left[H_{2,1}'\beta_{2,1}^*\right]_+\right)\\ &= \frac{e^{\alpha_n H_{2,1}'\beta_{2,1}^*}H_{2,1}'}{1+e^{\alpha_n H_{2,1}'\beta_{2,1}^*}}\sqrt{n}\left(\hat{\beta}_{2,1}-\beta_{2,1}^*\right) + o_P\left(\sqrt{n}\left(\hat{\beta}_{2,1}-\beta_{2,1}^*\right)\right)\\ &\quad + \sqrt{n}f(\alpha_n, H_{2,1}'\beta_{2,1}^*),\end{aligned}$$

where the last equality follows by taking derivatives, and where

$$f(\alpha,x) = \frac{1}{\alpha}\log\{1+e^{\alpha x}\} - \max(x,0). \tag{1}$$

For the remainder term, note that Lemma B.6 shows the consistency of $\hat{\beta}_{2,1}$, and that the expectation of the Hessian of $\frac{1}{\alpha_n}\log\{1+e^{\alpha_n H'\beta}\}$ is bounded by Assumption (A1). Hence, by applying Lemma B.5 to the matrix $\hat{\Sigma}_1$ that appears in the remainder term, we conclude that

$$\sqrt{n}(\hat{\beta}_1-\beta_1^*) = \mathbb{S}_n + \mathbb{T}_n + \mathbb{W}_n + o_P(1), \tag{2}$$

where

$$\begin{aligned}\mathbb{T}_n &= \hat{\Sigma}_1^{-1}\mathbb{P}_n B_1\left[\frac{e^{\alpha_n H_{2,1}'\beta_{2,1}^*}}{1+e^{\alpha_n H_{2,1}'\beta_{2,1}^*}}H_{2,1}'\sqrt{n}\left(\hat{\beta}_{2,1}-\beta_{2,1}^*\right)\right],\\ \mathbb{W}_n &= \hat{\Sigma}_1^{-1}\sqrt{n}\mathbb{P}_n B_1 f\left(\alpha_n, H_{2,1}'\beta_{2,1}^*\right).\end{aligned}$$

Recall that by assumption, α_n and thus fact that

$$\frac{e^{\alpha_n x}}{1+e^{\alpha_n x}} \to \begin{cases} 1 & x>0\\ \frac{1}{2} & x=0\\ 0 & x<0\end{cases},$$

we obtain that for a given $h_{2,1}$,

$$\frac{e^{\alpha_n h_{2,1}'\beta_{2,1}^*}}{1+e^{\alpha_n h_{2,1}'\beta_{2,1}^*}} \to \mathbf{1}\{h_{2,1}'\beta_{2,1}^*>0\} + \frac{1}{2}\mathbf{1}\{h_{2,1}'\beta_{2,1}^*=0\}.$$

Define the function $w: D_{p_1}\times l^\infty(\mathcal{F})\times\mathbb{R}^{p_{21}}\times[0,1]\mapsto\mathbb{R}^{p_1}$ by $w(\Sigma,\mu,\nu,a) = \Sigma^{-1}\mu(g(\nu,B_1,H_{2,1},a))$, where

$$g(\nu,b_1,h_{2,1},a) = \begin{cases} b_1\left[\frac{e^{h_{2,1}'\beta_{2,1}^*/a}}{1+e^{h_{2,1}'\beta_{2,1}^*/a}}h_{2,1}'\nu\right], & a>0\\ b_1\left[(\mathbf{1}\{h_{2,1}'\beta_{2,1}^*>0\}+\frac{1}{2}\mathbf{1}\{h_{2,1}'\beta_{2,1}^*=0\})h_{2,1}'\nu\right], & a=0\end{cases}$$

and where $\mathcal{F} = \{g(\nu, b_1, h_{2,1}, a), \|\nu\| \leq K\}$. Using the same arguments as those used in Lemma B.11, one can show that w is continuous at $(\Sigma_{1,\infty}, P, \mathbb{R}^{p_{21}}, 0)$. Thus, using the continuous mapping theorem, it can be shown that $\mathbb{S}_n + \mathbb{T}_n$ weakly converges to

$$\Sigma_{1,\infty}^{-1}\left[\mathbb{G}_\infty\left(B_1(H_{2,0}'\beta_{2,0}^* + \left[H_{2,1}'\beta_{2,1}^*\right]_+ - B_1'\beta_1^*)\right)\right]$$
$$+ \Sigma_{1,\infty}^{-1}\left[PB_1\left(H_{2,0}'\mathbb{Z}_{\infty,0} + H_{2,1}'\mathbb{Z}_{\infty,1}\big(\mathbf{1}\{H_{2,1}'\beta_{2,1}^* > 0\} + \frac{1}{2}\mathbf{1}\{H_{2,1}'\beta_{2,1}^* = 0\}\big)\right)\right]$$

where

$$\left(\mathbb{Z}_{\infty,0}', \mathbb{Z}_{\infty,1}'\right)' = \Sigma_{2,\infty}^{-1}\mathbb{G}_\infty\left[B_2(Y - B_2'\beta_2^*)\right].$$

We now discuss $\mathbb{W}_n$, the third term in (3). By Lemma 1(i) below, when $\sqrt{n}/\alpha_n \to 0$,

$$\left\|\hat{\Sigma}_1^{-1}\mathbb{W}_n\right\| \leq \sqrt{n}\mathbb{P}_n\left\|\hat{\Sigma}_1^{-1}\right\|\|B_1\|f(\alpha_n, H_{2,1}'\beta_{2,1}^*) \leq \frac{\sqrt{n}\log 2}{\alpha_n}\mathbb{P}_n\left\|\hat{\Sigma}_1^{-1}\right\|\|B_1\| \to 0,$$

which proves (i).

For (ii), let $\delta_n \to 0$ such that $\alpha_n\delta_n \to \infty$. Write

$$\begin{aligned}
\hat{\Sigma}_1\mathbb{W}_n &= \sqrt{n}\mathbb{P}_n\left(B_1 f\left(\alpha_n, H_{2,1}'\beta_{2,1}^*\right)\mathbf{1}\{|H_{2,1}'\beta_{2,1}^*| > \delta_n\}\right) \\
&\quad + \sqrt{n}\mathbb{P}_n\left(B_1 f\left(\alpha_n, H_{2,1}'\beta_{2,1}^*\right)\mathbf{1}\{|H_{2,1}'\beta_{2,1}^*| \leq \delta_n\}\right) \\
&= \sqrt{n}\mathbb{P}_n\left(B_1 f\left(\alpha_n, H_{2,1}'\beta_{2,1}^*\right)\mathbf{1}\{|H_{2,1}'\beta_{2,1}^*| > \delta_n\}\right) \\
&\quad + \sqrt{n}\mathbb{P}_n B_1 f\left(\alpha_n, H_{2,1}'\beta_{2,1}^*\right)\left(\mathbf{1}\{|H_{2,1}'\beta_{2,1}^*| \leq \delta_n\} - \mathbf{1}\{|H_{2,1}'\beta_{2,1}^*| = 0\}\right) \\
&\quad + \sqrt{n}\mathbb{P}_n\left(B_1 f\left(\alpha_n, H_{2,1}'\beta_{2,1}^*\right)\mathbf{1}\{|H_{2,1}'\beta_{2,1}^*| = 0\}\right) \\
&\equiv A_n + B_n + C_n.
\end{aligned}$$

Note that by Lemma 1(vi),

$$\|A_n\| \leq \mathbb{P}_n\|B_1\|\frac{\sqrt{n}}{\alpha_n}e^{-\alpha_n\delta_n} \xrightarrow{\mathrm{P}} 0.$$

Let $p(\delta) = P\left(\mathbf{1}\{|h_{2,1}'\beta_{2,1}^*| \leq \delta\} - \mathbf{1}\{|h_{2,1}'\beta_{2,1}^*| = 0\}\right)$, and note that $p(0) = 0$ and thus $p(\delta) \to 0$ as $\delta \to 0$. Hence,

$$\|B_n\| \leq \mathbb{P}_n\|B_1\|\frac{\sqrt{n}}{\alpha_n}\log(2)\|\left(\mathbf{1}\{|H_{2,1}'\beta_{2,1}^*| \leq \delta_n\} - \mathbf{1}\{|H_{2,1}'\beta_{2,1}^*| = 0\}\right)\| \xrightarrow{\mathrm{P}} 0.$$

Summarizing, we obtain that

$$\mathbb{W}_n \xrightarrow{\mathrm{P}} a_\infty\log(2)\Sigma_{1,\infty}^{-1}PB_1\mathbf{1}\{|H_{2,1}'\beta_{2,1}^*| = 0\},$$

which proves (ii). □

Sketch of proof of Theorem 2. Using the same arguments that lead to Eq. 2 in Laber et al., we have

$$\sqrt{n}(\hat{\beta}_1 - \beta^*_{1,n}) = \hat{\Sigma}_1^{-1}\mathbb{P}_n B_1 \left(\breve{Y} - B_1'\beta^*_{1,n}\right) = \mathbb{S}_n + \hat{\Sigma}_1^{-1}\mathbb{P}_n B_1 \mathbb{U}_n,$$

where $\mathbb{S}_n$ is smooth and asymptotically normal, and

$$\begin{aligned}\mathbb{U}_n = &\frac{e^{\alpha_n H_{2,1}'\beta^*_{2,1,n}} H_{2,1}'}{1+e^{\alpha_n H_{2,1}'\beta^*_{2,1,n}}}\sqrt{n}\left(\hat{\beta}_{2,1} - \beta^*_{2,1,n}\right) + o_P\left(\sqrt{n}\left(\hat{\beta}_{2,1} - \beta^*_{2,1,n}\right)\right)\\ &+ \sqrt{n}\left(\frac{1}{\alpha_n}\log\left\{1 + e^{\alpha_n H_{2,1}'\beta^*_{2,1,n}}\right\} - \left[H_{2,1}'\beta^*_{2,1,n}\right]_+\right).\end{aligned}$$

Similarly to the proof of Theorem 1, we have

$$\sqrt{n}(\hat{\beta}_1 - \beta^*_{1,n}) = \mathbb{S}_n + \mathbb{T}_n + \mathbb{W}_n + o_P(1), \tag{3}$$

where

$$\begin{aligned}\mathbb{T}_n &= \hat{\Sigma}_1^{-1}\mathbb{P}_n B_1\left[\frac{e^{\alpha_n H_{2,1}'\beta^*_{2,1,n}}}{1+e^{\alpha_n H_{2,1}'\beta^*_{2,1,n}}} H_{2,1}'\sqrt{n}\left(\hat{\beta}_{2,1} - \beta^*_{2,1}\right)\right],\\ \mathbb{W}_n &= \hat{\Sigma}_1^{-1}\sqrt{n}\mathbb{P}_n B_1 f\left(\alpha_n, H_{2,1}'\beta^*_{2,1,n}\right).\end{aligned}$$

We obtain that for a given $h_{2,1}$, and $\beta^*_{2,1,n} = \beta^*_{2,1} + \frac{s}{\sqrt{n}} + o(1/\sqrt{n})$, and $\alpha_n/\sqrt{n} \to a_\infty^{-1}$,

$$\frac{e^{\alpha_n h_{2,1}'\beta^*_{2,1,n}}}{1+e^{\alpha_n h_{2,1}'\beta^*_{2,1,n}}} \to \mathbf{1}\left\{h_{2,1}'\beta^*_{2,1} > 0\right\} + \left[a_\infty^{-1}h_{2,1}'s\right]_+ \mathbf{1}\left\{h_{2,1}'\beta^*_{2,1} = 0\right\}.$$

Using the same arguments given in the proof of Theorem 4.1 of Laber et al. (see also proof of Theorem 1 above) it can be shown that

$$\begin{aligned}\mathbb{S}_n + \mathbb{T}_n \rightsquigarrow \Sigma_{1,\infty}^{-1}\Big\{&\mathbb{G}_\infty\left(B_1(H_{2,0}'\beta^*_{2,0} + \left[H_{2,1}'\beta^*_{2,1}\right]_+ - B_1'\beta^*_1)\right)\\ &+ PB_1\left(H_{2,0}'\mathbb{Z}_{\infty,0} + H_{2,1}'\mathbb{Z}_{\infty,1}\mathbf{1}\left\{H_{2,1}'\beta^*_{2,1} > 0\right\}\right)\\ &+ PB_1\left(\left[a_\infty^{-1}H_{2,1}'s\right]_+\mathbf{1}\left\{H_{2,1}'\beta^*_{2,1} = 0\right\}\right)\Big\},\end{aligned}$$

where

$$\left(\mathbb{Z}_{\infty,0}', \mathbb{Z}_{\infty,1}'\right)' = \Sigma_{2,\infty}^{-1}\mathbb{G}_\infty\left[B_2(Y - B_2'\beta_2^*)\right].$$

Note that $\hat{\Sigma}_1\mathbb{W}_n$ can be written as

$$\begin{aligned}&\sqrt{n}\mathbb{P}_n\left(B_1 f\left(\alpha_n, H_{2,1}'\beta^*_{2,1,n}\right)\mathbf{1}\left\{|H_{2,1}'\beta^*_{2,1}| > \delta_n, |\alpha_n^{-1}H_{2,1}'s| > \delta_n/2\right\}\right)\\ &+ \sqrt{n}\mathbb{P}_n\left(B_1 f\left(\alpha_n, H_{2,1}'\beta^*_{2,1,n}\right)\mathbf{1}\left\{|H_{2,1}'\beta^*_{2,1}| > \delta_n, |\alpha_n^{-1}H_{2,1}'s| \le \delta_n/2\right\}\right)\\ &+ \sqrt{n}\mathbb{P}_n\left(B_1 f\left(\alpha_n, H_{2,1}'\beta^*_{2,1,n}\right)\left(\mathbf{1}\left\{|H_{2,1}'\beta^*_{2,1}| \le \delta_n\right\} - \mathbf{1}\left\{|H_{2,1}'\beta^*_{2,1}| = 0\right\}\right)\right)\end{aligned}$$

$$+\sqrt{n}\mathbb{P}_n\left(B_1 f\left(\alpha_n, H'_{2,1}\beta^*_{2,1,n}\right)\mathbf{1}\{|H'_{2,1}\beta^*_{2,1}|=0\}\right)$$
$$\equiv A_n + B_n + C_n + D_n.$$

The first three terms can be bounded as follows:

$$\|A_n + B_n + C_n\|$$
$$\le \mathbb{P}_n\|B\|\frac{\sqrt{n}}{\alpha_n}\log(2)\|\mathbf{1}\{|H'_{2,1}s|>\alpha_n\delta_n/2\}\| + \mathbb{P}_n\|B\|\frac{\sqrt{n}}{\alpha_n}e^{-\alpha_n\delta_n/2}$$
$$+\mathbb{P}_n\|B\|\frac{\sqrt{n}}{\alpha_n}\log(2)\left\|\mathbf{1}\{|H'_{2,1}\beta^*_{2,1}|\le\delta_n\}-\mathbf{1}\{|H'_{2,1}\beta^*_{2,1}|=0\}\right\| \xrightarrow{\mathrm{P}} 0.$$

Thus the limiting distribution of $\mathbb{W}_n$ depends on that of D_n. We have that D_n equals

$$\mathbb{P}_n B_1\left(\frac{\sqrt{n}}{\alpha_n}\log\left\{1+e^{\frac{\alpha_n}{\sqrt{n}}H'_{2,1}s+o(1/\sqrt{n})}\right\}-\left[H'_{2,1}s+o(1)\right]_+\right)\mathbf{1}\{|H'_{2,1}\beta^*_{2,1}|=0\}$$
$$\xrightarrow{\mathrm{P}} PB_1\left(a_\infty\log\left\{1+e^{a_\infty^{-1}H'_{2,1}s}\right\}-\left[H'_{2,1}s\right]_+\right)\mathbf{1}\{|H'_{2,1}\beta^*_{2,1}|=0\},$$

which concludes the proof. □

Proof of Corollary 1. We prove only the second assertion, the first can be proved similarly. Noting that both $\mathbb{S}_\infty$ and $\mathbb{T}_\infty$ have mean zero, it is enough to bound the bias of

$$\Sigma^{-1}_{1,\infty}P(\mathbb{W}_\infty)$$
$$=\Sigma^{-1}_{1,\infty}PB_1\left(a_\infty\log\left\{1+e^{a_\infty^{-1}H'_{2,1}s}\right\}-\left[H'_{2,1}s\right]_+\right)\mathbf{1}\{|H'_{2,1}\beta^*_{2,1}|=0\}.$$

Using Lemma 1(i) with $\alpha = a_\infty^{-1}$ and $x = H'_{2,1}s$, we obtain that the bias is bounded by

$$a_\infty\log(2)\|\Sigma^{-1}_{1,\infty}\|P\|B\|\mathbf{1}\{|H'_{2,1}\beta^*_{2,1}|=0\}.$$ □

The following lemma is needed for the proofs of Theorems 1–2:

Lemma 1. *Let* $f(\alpha,x)=\frac{1}{\alpha}\log\{1+e^{\alpha x}\}-\max(x,0)$. *Then*

(i) $0<f(\alpha,x)\le\log(2)/\alpha$.
(ii) $\operatorname{argmax}_x f(\alpha,x)=0$ *and* $f(\alpha,0)=\log(2)/\alpha$.
(iii) $f(\alpha,x)\to 0$ *as* $\alpha\to\infty$.
(iv) Fix $\delta>0$. *Then* $\max_{x\in(-\infty,-\delta]\cup[\delta,\infty)}f(\alpha,x)=\frac{e^{-\alpha\delta}}{\alpha}$.

The proof is technical and therefore omitted.

References

Goldberg, Y., Song, R., and Kosorok, M. R. (2013), *From Probability to Statistics and Back: High-Dimensional Models and Processes – A Festschrift in Honor of Jon A. Wellner*, Institute of Mathematical Statistics, vol. 9 of *IMS Collections*, chap. Adaptive Q-Learning, pp. 150–162. MR3186754

Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E., and Murphy, S. A. (2014), Dynamic treatment regimes: Technical challenges and applications. *Electron. J. Statist.* 8 1225–1272.

Moodie, E. E. M. and Richardson, T. S. (2010), Estimating optimal dynamic regimes: Correcting bias under the null, *Scandinavian Journal of Statistics*, 37, 126–146. MR2675943

Song, R., Wang, W., Zeng, D., and Kosorok, M. R. (2011), Penalized Q-learning for dynamic treatment regimes, *To appear in Statistical Sinica*.

Zhang, C. H. and Zhang, S. S. (2014), Confidence intervals for low dimensional parameters in high dimensional linear models, *Journal of the Royal Statistical Society: Series B*, 76, 217–242. MR3153940