ADAPTIVE CONFIDENCE INTERVALS FOR NON-SMOOTH FUNCTIONALS

by

Eric B. Laber

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2011

Doctoral Committee:

Professor Susan A. Murphy, Chair
Professor Satinder Singh Baveja
Associate Professor Kerby A.Shedden
Associate Professor Ji Zhu

To my father

ACKNOWLEDGEMENTS

This thesis would not have been possible without the patience and generous support of my friends, family, colleagues, and professors. I cannot hope that a few kind words would begin to repay what I feel I owe to all of the people that have helped me along the way. I only wish that they see my thanks here as a token of the gratitude I will never be able to fully express.

First, I would like thank my advisor Professor Susan Murphy. She taught me that understanding is of the highest priority, and that thoughtful, honest research always takes precedence over publication count. Her guidance, criticism, patience, and generosity hold sway in everything I have done and will do as a researcher. Her help has made everything I do better.

I also owe a special debt to the members of my committee: Professors Kerby Shedden, Satinder Singh, and Ji Zhu for their suggestions and comments over the years.

I would also like to thank the faculty and staff in the Statistics Department and at the Institute for Social Research. In particular, Mary Ann King, Lu Ann Custer, Amy Rundquist, and Rhonda Moats, without their help I would have: i) been kicked out of school, ii) missed any number of my flights, iii) never had a prelim, and iv) starved the statistics working groups. I also would like to make special mention of Kerby Shedden (rapper name KSHED). Our many conversations have had tremendous influence on the way I think about life and statistics, and our many bike rides have helped to keep me a little less fat.

I would like to thank my classmates for their help over the years. I could not have hoped for a better cohort. In particular Joel Vaughn and Yang Yang have seen every talk I have ever given and could probably defend this thesis in a pinch. You two are the best.

I would also like to thank the people who have come through Professor Susan Murphy's lab over the years: Danny Almirall, Anne Bu, Bibhas Chakraborty, Kun Deng, Lacey Gunter, Zhiguo Li, Dan Lizotte, Min Qian, Lu Xie, and Peng Zhang.

Many persons outside the University of Michigan are also owed my gratitude: Cam Bottolfson, Daryl Fisher, Mark Hansen, Dusty Jensen, Larry Kahlow, Diane Lambert, Doc Newman, Natesh Pillai, and Matt Zabka.

Most importantly, I would like to offer thanks to my family. I would like to thank my father for his endless support, it is his interest in science that makes me wish statisticians wore lab coats. I would like to thank my siblings and their families for their love and support. I also offer thanks to the Nair family for taking me in as one of their own. I thank Professor Vijay Nair for taking on the role of a mentor and a father figure. Last, I offer thanks to Sheela Nair, without whom, all of this would mean nothing.

TABLE OF CONTENTS

DEDICATION ii

ACKNOWLEDGEMENTS iii

LIST OF FIGURES vii

LIST OF TABLES ix

CHAPTER

I. Introduction 1

1.1 Non-smoothness and non-regularity 2
1.2 Outline 5

II. Adaptive confidence intervals for the test error in classification 8

2.1 Introduction 9
2.2 Motivation 11
2.3 Adaptive confidence interval 16
2.3.1 Construction of the ACI 16
2.3.2 Properties of the ACI 18
2.3.3 Local Alternatives 22
2.3.4 Choice of Tuning Parameter a_n 24
2.4 Computation 25
2.5 Empirical study 27
2.5.1 Competing methods 29
2.5.2 Results 31
2.6 Discussion 33
2.7 Appendix I: Additional experimental results 36
2.8 Appendix II: Proofs 39

III. Statistical inference for dynamic treatment regimes 45

3.1 Two stages of binary treatment 49
3.2 Extending the ACI to many treatments 60
3.3 Empirical Study 65
3.3.1 The choice of λ_n 66
3.3.2 An Evaluation of the ACI 68
3.4 Analysis of the ADHD study 72
3.5 Discussion 79
3.6 Appendix I: Proofs for the ACI for two stages and two treatments per stage 80
3.7 Appendix II: Extension of the ACI to many stages and many treatments 116
3.7.1 Q-Learning in the general case 116
3.7.2 ACI in the general case 118
3.7.3 Properties of the ACI in the general case 123
3.8 Appendix III: Proofs for the ACI in more than two stages and more than two treatments 126
3.9 Appendix IV: Bias reduction for non-regular problems 139
3.10 Appendix V: Additional empirical results 143
3.10.1 Varying dataset size 143
3.10.2 Models with ternary actions 156
3.10.3 Models with three stages 170

IV. Future work 184

4.1 Inference after model selection via the Lasso 184
4.1.1 Non-regularity of the Lasso 186
4.1.2 Adaptive confidence intervals for the Lasso 189
4.1.3 Computation 192
4.1.4 Conclusions 193
4.2 Adaptive confidence intervals for the value of a learned DTR 194

BIBLIOGRAPHY 197

LIST OF FIGURES

Figure

2.1 Left: Coverage of centered percentile bootstrap and normal approximations for constructing confidence sets for $\tau(\hat{c})$. Right: Coverage of centered percentile bootstrap with smoothed target $\tau_{smoothed}(\hat{c}) \triangleq P(1 + exp(aY\hat{c}(X)))^{-1}$ for varying values of a; a value of $a = \infty$ corresponds to $\tau(\hat{c})$. Results are based on 1000 Monte Carlo iterations, target coverage is .950. The performance of the ACI on this example can found in Section 5 under the example labeled "quad." 15

2.2 Relaxation of the indicator functions. Left panel: indicator function $1_{yx^\intercal u<0}$ replaced with convex surrogate $(1 - yx^\intercal u)_+$. Right panel: indicator function $-1_{yx^\intercal u<0}$ replaced with convex surrogate $(1 + yx^\intercal u)_+ - 1$. 27

2.3 Computation time for MIP and relaxed construction of ACI using the ThreePt data set and squared error loss. 29

3.1 Left: The bias of $\hat{\theta}_{\text{soft}}$ as a function of effect size $\mu_1 - \mu_2$ and tuning parameter λ. Reducing the bias at $\mu_1 - \mu_2 = 0$ requires increasing λ which is seen to dramatically inflate bias elsewhere. Right: The MSE of $\hat{\theta}_{\text{soft}}$ as a function of effect size $\mu_1 - \mu_2$ and tuning parameter λ. Attempting to reduce the bias at $\mu_1 - \mu_2 = 0$ results in a modest reduction in MSE at $\mu_1 - \mu_2 = 0$ but inflates the MSE significantly elsewhere. 141

3.2 Left: The bias of $\hat{\theta}_{\text{hard}}$ as a function of effect size $\mu_1 - \mu_2$ and tuning parameter λ. Reducing the bias at $\mu_1 - \mu_2 = 0$ requires increasing λ which is seen to dramatically inflate bias elsewhere. Right: The MSE of $\hat{\theta}_{\text{hard}}$ as a function of effect size $\mu_1 - \mu_2$ and tuning parameter λ. Attempting to reduce the bias at $\mu_1 - \mu_2 = 0$ results in a modest reduction in MSE at $\mu_1 - \mu_2 = 0$ but inflates the MSE significantly elsewhere. 142

4.1 Estimated coverage of the residual bootstrap (Residual BS) and the adaptive confidence interval (ACI). Target coverage is 95%. Estimates are based on 1000 Monte Carlo iterations and 1000 bootstrap resamples. 189

LIST OF TABLES

Table

2.1 Comparison of MIP and relaxed versions of the ACI. For each data set the table was constructed using 1000 training sets each with 1000 bootstrap iterations for a total of 1,000,000 computations of the optimization problem given in (2.12). 30

2.2 Test data sets used to evaluate confidence interval performance. The last three columns record the average test error for a linear classifier trained using a training set of size $n = 100$ and loss function: squared error loss (SE), binomial deviance (BD), and ridged hinge loss (SVM). 32

2.3 Coverage comparison between ACI, Yang's CV and Jiang's $BCCVP-BR$ for squared error loss, target coverage is .950. Coverage is starred if observed coverage is significantly different from .950 at .01 level. 33

2.4 Comparison of interval width between ACI, Yang's CV and Jiang's $BCCVP - BR$ for squared error loss. Smallest observed width is starred. Examples where at least the nominal coverage was not attained are omitted. 33

2.5 Comparison of computation time (in seconds) between ACI, Yang's CV and Jiang's $BCCVP - BR$ for squared error loss. Examples where at least the nominal coverage was not attained are omitted. . 34

2.6 Coverage comparison between ACI, Yang's CV and Jiang's $BCCVP-BR$ for binomial deviance loss, target coverage is .950. 36

2.7 Comparison of interval width between ACI, Yang's CV and Jiang's $BCCVP - BR$ for binomial deviance loss. Entries with the smallest width are starred. Examples where at least the nominal coverage was not attained are omitted. 36

2.8 Coverage comparison between ACI, Yang's CV and Jiang's $BCCVP-BR$ for ridged (L_2 penalized) hinge loss, target coverage is .950. . . 37

2.9 Comparison of interval width between ACI, Yang's CV and Jiang's $BCCVP-BR$ for ridged (L_2 penalized) hinge loss. Entries with the smallest width are starred. Examples where at least the nominal coverage was not attained are omitted. 37

2.10 Coverage and width of ACI using squared error loss and a training set size of $n = 30$. Coverage and width appears to be stable across a range of reasonable values for tuning parameter γ. 38

2.11 Coverage of Centered Percentile Bootstrap (CPB) using three loss functions (squared loss, logistic loss, and hinge loss) and three sample sizes ($n = 30$, $n = 100$, and $n = 250$). Target coverage is .950, coverage is starred if significantly different from target. 38

3.1 Parameters indexing the example models. 67

3.2 Monte Carlo estimates of coverage probabilities for the ACI methods at the 95% nominal level. Here, $\beta_{1,1,1}$ denotes the main effect of treatment and $\beta_{1,0,1}$ denotes the intercept. Estimates are constructed using 10000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Examples are designated NR = non-regular, NNR = near-non-regular, R = regular. 69

3.3 Monte Carlo estimates of mean width of the ACI method at the 95% nominal level. Here, $\beta_{1,1,1}$ denotes the main effect of treatment and $\beta_{1,0,1}$ denotes the intercept. Estimates are constructed using 10000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Models have two treatments at each of two stages. Examples are designated NR = non-regular, NNR = near-non-regular, R = regular. 69

3.4 Monte Carlo estimates of coverage probabilities of confidence intervals for the coefficient of the treatment variable, $\beta^*_{1,1,1}$ at the 95% nominal level. Estimates are constructed using 1000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. There is no ST method when there are three treatments at Stage 2. Examples are designated NR = non-regular, NNR = near-non-regular, R = regular. 73

3.5 Monte Carlo estimates of mean width of confidence intervals for the coefficient of the treatment variable, $\beta^*_{1,1,1}$ at the 95% nominal level. Estimates are constructed using 1000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Models have two treatments at each of two stages. Widths with corresponding coverage significantly below nominal are marked with $*$. There is no ST method when there are three treatments at Stage 2. Examples are designated NR = non-regular, NNR = near-non-regular, R = regular. 73

3.6 Monte Carlo estimates of coverage probabilities of confidence intervals for the coefficient of the intercept, $\beta^*_{1,0,1}$ at the 95% nominal level. Estimates are constructed using 1000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Examples are designated NR = non-regular, NNR = near-non-regular, R = regular. 74

3.7 Monte Carlo estimates of mean width of confidence intervals for the coefficient of the intercept, $\beta^*_{1,0,1}$ at the 95% nominal level. Estimates are constructed using 1000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Models have two treatments at each of two stages. Widths with corresponding coverage significantly below nominal are marked with $*$. Examples are designated NR = non-regular, NNR = near-non-regular, R = regular. 74

3.8 Components of a single trajectory in the ADHD study. 75

3.9 Least squares coefficients and 90% interval estimates for second stage regression. 76

3.10 Least squares coefficients and 90% ACI interval estimates for first stage regression. 77

3.11 Confidence intervals for the predicted difference in mean response across treatments across different patient histories at the 90% level. Confidence intervals that do not contain zero are deemed as having insufficient evidence for recommending a unique best treatment. . . 79

3.12 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,0}$ (intercept term) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 144

3.13 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,0}$ (intercept term) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 145

3.14 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,1}$ (main effect of history) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 146

3.15 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,1}$ (main effect of history) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 147

3.16 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}$ (interaction between history and treatment) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 148

3.17 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}$ (interaction between history and treatment) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 149

3.18 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,1}$ (main effect of treatment) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 150

3.19 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,1}$ (main effect of treatment) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 151

3.20 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}+\beta^*_{1,1,1}$ (effect of action for history = 1) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 152

3.21 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}+\beta^*_{1,1,1}$ (effect of action for history = 1) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 153

3.22 Monte Carlo estimates of coverage probabilities of confidence intervals for $-\beta^*_{1,1,0}+\beta^*_{1,1,1}$ (effect of action for history = -1) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 154

3.23 Monte Carlo estimates of coverage probabilities of confidence intervals for $-\beta^*_{1,1,0} + \beta^*_{1,1,1}$ (effect of action for history = -1) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 155

3.24 Parameters indexing the example models. 157

3.25 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,0}$ (intercept term) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 158

3.26 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,0}$ (intercept term) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 159

3.27 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,1}$ (main effect of history) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 160

3.28 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,1}$ (main effect of history) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 161

3.29 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}$ (interaction between history and treatment) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 162

3.30 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}$ (interaction between history and treatment) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 163

3.31 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,1}$ (main effect of treatment) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 164

3.32 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,1}$ (main effect of treatment) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 165

3.33 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}+\beta^*_{1,1,1}$ (effect of action for history = 1) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 166

3.34 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}+\beta^*_{1,1,1}$ (effect of action for history = 1) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 167

3.35 Monte Carlo estimates of coverage probabilities of confidence intervals for $-\beta^*_{1,1,0}+\beta^*_{1,1,1}$ (effect of action for history = -1) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 168

3.36 Monte Carlo estimates of coverage probabilities of confidence intervals for $-\beta^*_{1,1,0} + \beta^*_{1,1,1}$ (effect of action for history = -1) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 169

3.37 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,0}$ (intercept term) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 172

3.38 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,0}$ (intercept term) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 173

3.39 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,1}$ (main effect of history) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 174

3.40 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,1}$ (main effect of history) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 175

3.41 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}$ (interaction between history and treatment) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 176

3.42 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}$ (interaction between history and treatment) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 177

3.43 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,1}$ (main effect of treatment) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 178

3.44 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,1}$ (main effect of treatment) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 179

3.45 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}+\beta^*_{1,1,1}$ (effect of action for history = 1) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 180

3.46 Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}+\beta^*_{1,1,1}$ (effect of action for history = 1) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 181

3.47 Monte Carlo estimates of coverage probabilities of confidence intervals for $-\beta^*_{1,1,0} + \beta^*_{1,1,1}$ (effect of action for history = -1) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 182

3.48 Monte Carlo estimates of coverage probabilities of confidence intervals for $-\beta^*_{1,1,0} + \beta^*_{1,1,1}$ (effect of action for history = -1) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular. 183

CHAPTER I

Introduction

Many quantities of interest in modern statistical analysis are non-smooth functionals of the underlying generative distribution, the observed data, or both. Examples include the test error of a learned classifier, parameters indexing an estimated optimal dynamic treatment regime, and the coefficients in a regression model after model selection has been performed. This lack of smoothness can lead to non-regular asymptotics under many 'real-life' scenarios and thus invalidate standard statistical procedures like the bootstrap and series approximations. Statistical procedures that either ignore or assume away this non-regularity can perform quite poorly, especially in small samples.

The aim of this dissertation is (i) to illustrate the impact that non-regularity can have on the performance of statistical inference procedures, especially in small samples, and (ii) to develop tools for conducting theoretically valid statistical inference for non-smooth functionals. In particular, we aim to develop confidence intervals that deliver asymptotically correct coverage under both fixed and local alternatives. To construct confidence intervals we first derive smooth, data-dependent, upper and lower bounds on the functional of interest and then approximate the distribution of the bounds using standard techniques. We then use estimated distributional features, such as the quantiles, to make inference for the original non-smooth functional. We leverage the smoothness of the bounds to obtain consistent inference under both fixed and local alternatives. This consistency is instrumental in ensuring good performance in both in both large and small samples. An important feature of these bounds is that they are adaptive to the underlying non-smoothness of the functional. That is, they are asymptotically tight in the case when the generative distribution happens to induce sufficient smoothness.

1.1 Non-smoothness and non-regularity

Non-smoothness and non-regularity are intimately connected. In particular, non-smooth functions of regular estimators become non-regular estimators. Non-regular estimators arising in this way comprise a major proportion of this thesis. Another way in which non-regular estimators arise are as the minimizer (maximizer) of either a constrained or non-smooth objective function. The Lasso estimator (Tibshirani 1996) is perhaps the best known example of this type. We briefly discuss the problem of inference for Lasso type estimators in the future work section of this thesis. In the remainder of this section we define formally the concept of a non-regular estimator and show how non-smoothness can induce non-regularity using a toy example.

Suppose we observe $X_1, X_2, \ldots, X_n$ drawn independently from some fixed but unknown distribution P. Furthermore, suppose we are interested in estimating some feature of P, say $\theta(P) \in \Theta \subseteq \mathbb{R}^p$, using $X_1, X_2, \ldots, X_n$. Let $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ denote our estimator of $\theta(P)$. Assume $\sqrt{n}(\hat{\theta} - \theta(P)) \rightsquigarrow L$, where L is tight and non-degenerate, and "$\rightsquigarrow$" denotes weak convergence. For each n, consider the triangular array $X_{n,1}, X_{n,2}, \ldots, X_{n,n}$ drawn *iid* from distribution P_n. It is assumed that P_n is a sequence of contiguous alternatives to P in the sense of van der Vaart and Wellner

(1996). That is,

$$\int \left[\sqrt{n} \left(dP_n^{1/2} - dP^{1/2} \right) - \frac{1}{2} h dP^{1/2} \right]^2 \to 0,$$

as n tends to ∞ for some measurable function h. Furthermore, assume that $\theta(P_n) = \theta(P) + u/\sqrt{n}$. If $\sqrt{n}(\hat{\theta}(X_{n,1}, X_{n,2}, \ldots, X_{n,n}) - \theta(P_n)) \rightsquigarrow L_u$ under P_n, and L_u is equal in law to L for all u, then we say that $\hat{\theta}$ is *regular*. On the other hand, if L_u depends on u, then the estimator $\hat{\theta}$ is said to be *non-regular*. Intuitively, non-regular estimators can be highly sensitive to small perturbations in the generative distribution. Consequently, standard asymptotic approximations based on such estimators can be inaccurate and lead to poor small sample performance.

To make concrete the notion of a regular estimator we consider a toy example. Suppose that $X_1, X_2, \ldots, X_n$ are *iid* univariate normal random variables with unknown mean θ an unit variance. The sample mean $\hat{\theta} \triangleq \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is a regular estimator of θ. To see this, for each n consider a triangular array $X_{n,1}, X_{n,2}, \ldots, X_{n,n}$ drawn *iid* from a univariate normal distribution with mean $\theta_n = \theta + u/\sqrt{n}$ and unit variance. Here, u is called the local parameter. Then, it is easy to check that $\sqrt{n}(\hat{\theta}(X_{n,1}, X_{n,2}, \ldots, X_{n,n}) - \theta_n) \rightsquigarrow Z$ where Z is a standard normal random variable. Since the limiting distribution does not depend on the local parameter u, we see that $\hat{\theta}$ is regular. On the other hand, consider the task of estimating $\kappa(\theta) \triangleq |\theta|$. The plug-in estimator $\kappa(\hat{\theta}) = |\bar{X}_n|$ is non-regular since under the local normal model

$$\sqrt{n}(\kappa(\hat{\theta}(X_{n,1}, X_{n,2}, \ldots, X_{n,n}) - \kappa(\theta_n)) \rightsquigarrow \begin{cases} |Z + u| - |u| & \text{if } \theta = 0, \\ Z & \text{if } \theta \neq 0, \end{cases}$$

where Z denotes a standard normal random variable. Since the local parameter is present in the above limiting distribution we see that $\kappa(\hat{\theta})$ is non-regular. More specifically, since the local parameter only appears in the limit when $\theta = 0$ we say that $\kappa(\hat{\theta})$ is non-regular at $\theta = 0$.

The preceding example showed that $\kappa(\hat{\theta})$ is non-regular at $\theta = 0$. This might lead one to think that this non-regularity is merely the consequence of a poor choice of estimator and that a more judicious choice will eliminate this non-regularity. On the contrary, a regular estimator of $\kappa(\theta)$ does not exist. More precisely, suppose that $\hat{\kappa} = \hat{\kappa}(X_1, X_2, \ldots, X_n)$ is any estimator of $\kappa(\theta)$. Furthermore, for each θ, suppose that if we consider data generated under the local process $\theta_n \triangleq \theta + u/\sqrt{n}$, then for each value of the local parameter u, $\hat{\kappa}$ satisfies

$$\sqrt{n}(\hat{\kappa}(X_{n,1}, X_{n,2}, \ldots, X_{n,n}) - \kappa(\theta + u/\sqrt{n})) \rightsquigarrow L_u(\theta),$$

where L_u is a tight (possibly degenerate) distribution, then $\hat{\kappa}$ *must* be non-regular at $\theta = 0$ (see van der Vaart 1991; Porter and Hirano and Porter 2009).

The preceding example shows a special case of a general phenomenon in which non-smooth functionals of the generative distribution frequently do not admit regular estimators (van der Vaart 1991; Hirano and Porter 2009). Regularity is critical for the validity of standard asymptotic approaches to inference like the bootstrap or series approximations (see Doss and Sethuraman; Shao 1994; Beran 1997). Thus, valid inference often requires specialized methodology, tailor-made to the individual problem at hand. This thesis attempts to take first steps toward a general framework for the construction of confidence sets for a large class of non-regular problems.

While each chapter contains its own individual literature review, we briefly mention an alternative approach to constructing confidence sets for non-regular problems that has received a good deal of attention in the literature: the “m-out-of-n” bootstrap (Bretagnolle 1983). The m-out-of-n bootstrap is a way of restoring large sample consistency in many non-regular problems. Unfortunately, for reasons mentioned below, the small sample performance of the m-out-of-n bootstrap can be poor in some settings.

Confidence intervals formed using the m-out-of-n bootstrap are formed in exactly the same way as when using the multinomial bootstrap, except that the size of the resampled data sets are taken to be m rather than the original data set size n. Viewing the resample size m as a function of data set size n, say $m = m(n)$, and choosing $m(n)$ to tend to ∞ as $n \to \infty$ and also satisfy $m(n) = o(n)$ leads to asymptoticaly valid confidence intervals for many non-regular problems (see Bretagnolle 1983; Beran and Srivastava 1985; Athreya 1987; Politis and Romano 1994; Shao 1994; Bickel et al. 1997). Intuitively, the m-out-of-n bootstrap works by letting n tend to infinity 'before' $m(n)$ does. In this way, asymptotic theory can proceed as if one were simply resampling from the true generative distribution rather than resampling the empirical measure (see Shao 1994 for a formal discussion and additional details). While the m-out-of-n bootstrap often provides valid asymptotics, its utility in small samples can be limited (see Dumbgen 1993; Beran 1997; Andrews 2000; Andrews and Guggenberger 2005abcd; Samworth 2003; Andrews 2008). In particular, small sample performance tends to be highly sensitive to the choice of m. Furthermore, the nature of the m-out-of-n bootstrap is asymptotic in both m and n and thus requires not only that the sample size n be relatively large, but also that m must be large as well.

1.2 Outline

Broadly speaking, this thesis offers a novel framework for constructing confidence intervals for non-regular functionals. These confidence intervals are formed by bounding the non-regular functional of interest between two smooth (and regular) functionals. The smoothness of these bounds can be leveraged to obtain useful asymptotic results.

In particular, confidence intervals based on these bounds are shown to be automatically adaptive to the amount of non-smoothness in the generative model in the sense that they provide asymptotically exact coverage under sufficiently smooth generative models, and are otherwise conservative. Furthermore, we establish consistency results under both fixed and local alternatives.

We term a confidence interval constructed using this bounding approach an Adaptive Confidence Interval (ACI) for the adaptivity properties mentioned above. This thesis introduces the ACI and its properties for the test error in classification and inference for parameters indexing an optimal dynamic treatment regime. An outline of the remainder of this thesis is as follows. Chapter II introduces the ACI for the test error in classification. In this chapter, we show that the test error is a non-smooth functional of the data and the underlying generative distribution. We also show that application of standard approaches like the bootstrap and normal approximations provide extremely poor small sample performance. The ACI is introduced and shown to be consistent under fixed and local alternatives. Furthermore, the ACI is shown to be exact if either (i) the model space is chosen correctly, or (ii) the generative model satisfies a margin condition. The ACI is also shown to compare favorably to competitors in a suite of empirical experiments.

Chapter III introduces the ACI for parameters indexing an estimated optimal dynamic treatment regime. The ACI is shown to be asymptotically correct coverage under both fixed and local alternatives. In addition, the ACI is developed for both an arbitrary number of treatments and an arbitrary number of stages of treatment. The ACI is shown to perform well in a number of empirical experiments and is applied to analyze data from the Adaptive Interventions for Children with ADHD study (Pelham and Fabiano 2008).

Chapter IV discusses two extensions of the ACI methodology. The first is inference for the parameters estimated via the Lasso. This is an important extension because the bounds used in the ACI must be defined implicitly through the penalized least squares criterion defining the Lasso. Thus, developing the ACI for the Lasso, makes significant strides toward a general framework for M-estimators. The second extension is for the value function in dynamic treatment regimes. This quantity can be seen as multistage analogue of the test error in classification. The value function is highly non-smooth and is also of great interest to decision makers.

CHAPTER II

Adaptive confidence intervals for the test error in classification

The estimated test error of a learned classifier is the most commonly reported measure of classifier performance. However, constructing a high quality point estimator of the test error has proved to be very difficult. Furthermore, common interval estimators (e.g. confidence intervals) are based on the point estimator of the test error and thus inherit all the difficulties associated with the point estimation problem. As a result, these confidence intervals do not reliably deliver nominal coverage. In contrast we directly construct the confidence interval by use of smooth data-dependent upper and lower bounds on the test error. We prove that for linear classifiers, the proposed confidence interval automatically adapts to the non-smoothness of the test error, is consistent under fixed and local alternatives, and does not require that the Bayes classifier be linear. Moreover, the method provides nominal coverage on a suite of test problems using a range of classification algorithms and sample sizes.

2.1 Introduction

In classification problems, we observe a training set of (feature, label) pairs, $\mathcal{T} = \{(X_i, Y_i)\}_{i=1}^n$. The goal is use this sample to construct a classifier, say $\hat{c}$, so that when presented with a new feature, X, $\hat{c}(X)$ will accurately predict the unobserved label, Y. Accurate prediction corresponds to small test error; recall that the test error is given by $\tau(\hat{c}) = P1_{\hat{c}(X) \neq Y}$ where $P1_{\hat{c}(X) \neq Y} = \int 1_{\hat{c}(x) \neq y} dP(x, y)$ denotes expectation over the distribution P of (X, Y) only, and not the distribution of the training set. The test error $\tau(\hat{c})$ is a functional of $\hat{c}$ and thus is a random quantity. For this reason $\tau(\hat{c})$ is sometimes referred to as the conditional test error (Efron 1997; Hastie et al. 2009; Chung and Han 2009). Estimation of the test error typically employs resampling. Most commonly, the leave-one-out or k-fold cross-validated test error is reported in practice. Bootstrap estimates of the test error were suggested by Efron (1983) and later refinements were given by Efron and Tibshirani (1995, 1997). There have been a number of simulation studies comparing these approaches; some references include (Efron 1983; Chernick et al. 1985; Kohavi 1995; Krzanowksi and Hand 1996). A nice survey of estimators is given by Schiavo and Hand (2000). However many have documented that estimators of the test error are plagued by bias and high variance across training sets (Zhang 1995; Isaakson 2008; Hastie et al. 2009) and consequently the test error is accepted to be a difficult quantity to estimate accurately. Two reasons for this problematic behavior are that some classification algorithms result in a $\hat{c}$ that is a non-smooth functional of the training set, and, even when $\hat{c}$ is a smooth functional of the training set, the test error is the expectation of a non-smooth function of $\hat{c}$.

An alternative to point estimation is interval estimation (e.g. a confidence interval). However, this approach has also been problematic likely because researchers have followed what we call the "point estimation paradigm": as a first step a point estimator of the test error is constructed, and as a second step, the distribution of this estimator is approximated. The problem with this approach is that a problematic point estimator of the test error makes the second step very difficult. The point estimation paradigm was employed by Efron and Tibshirani (1997) where the standard error of their smoothed leave-one-out estimator was approximated using the nonparametric delta method. Efron and Tibshirani noted that this approach would not work, however, for their more refined .632 (or .632+) estimators because of non-smoothness. Yang (2006) follows this paradigm as well, using a normal approximation to the repeated split cross-validation estimator. In practice, the point estimation paradigm is often applied by simply bootstrapping the estimator of the test error (see Jiang et al., 2008; Chung and Han 2009). These methods, while intuitive, lack theoretical justification.

We consider interval estimators for linear classifiers constructed from training sets in which the number of features is less than the training set size ($p << n$). As will be seen, even in this simple setting, natural approaches to constructing interval estimators for the test error can perform poorly. Instead of using the point estimation paradigm, we directly construct the confidence interval by use of smooth data-dependent upper and lower bounds on the test error. These bounds are sufficiently smooth so that their bootstrap distribution can be used to construct valid confidence intervals. Moreover, these bounds are adaptive in the sense that under certain settings exact coverage is delivered.

The outline of this paper is as follows. In Section 2 we illustrate the small sample problems that motivate the use of approximations in a non-regular asymptotic framework. Section 3 introduces the Adaptive Confidence Interval (ACI). The ACI is shown to be consistent under fixed and local alternatives. Section 4 addresses the computational issues involved in constructing the ACI. A computationally efficient (polynomial time) convex relaxation of the ACI is developed and shown to provide nearly identical results to exact computation. Section 5 provides a large experimental study of the ACI and several competitors. A variety of classifiers and sample sizes are considered on a suite of ten examples. The ACI is shown to provide correct coverage while being shorter in length than competing methods. Section 6 discusses a number of generalizations and directions for future research. Most proofs are left to the appendix.

2.2 Motivation

Throughout we assume that the training set is an *iid* sample $\mathcal{T} = \{(X_i, Y_i)\}_{i=1}^{n}$ drawn from some unknown joint distribution P. The features X are assumed to take values in $\mathbb{R}^p$ while the labels are coded $Y \in \{-1, 1\}$. To construct the linear classifier we fit a linear model $\hat{f}_{\mathcal{T}}(x) = x^{\intercal}\hat{\beta}_n$ by minimizing a convex criterion function. That is, we construct $\hat{\beta}_n \triangleq \arg\min_{\beta \in \mathbb{R}^p} \mathbb{P}_n L(X, Y, \beta)$ where $\mathbb{P}_n$ is the empirical measure and $L(X, Y, \beta)$ is a convex function of β (e.g., hinge loss with an L_2 penalty in the case of linear support vector machines). The classifier is the sign of the linear fit; that is, the predicted label y at input x is assigned according to $\hat{c}(x) = sign(x^{\intercal}\hat{\beta}_n)$ (define $sign(0) = 1$). Recall that the test error of the learned classifier is defined as

$$\tau(\hat{c}) \triangleq P1_{sign(X^{\intercal}\hat{\beta}_n) \neq Y} = P1_{YX^{\intercal}\hat{\beta}_n < 0},$$

where P denotes expectation with respect to X and Y.

As discussed in the introduction, the test error is a non-smooth functional of the training data. To see this and to gain a clearer understanding of the test error note

$$\tau(\hat{c}) = const + \int\limits_{\mathbb{R}^p} [2q(x) - 1]\, 1_{x^\intercal \hat{\beta}_n < 0} dP_X(x), \tag{2.1}$$

where $q(x) \triangleq P(Y = 1|X = x)$. Recall that $sign(2q(x) - 1)$ is the Bayes classifier. Then

$$Var\left(\tau(\hat{c})\right) = \mathbb{E}\left(\int\limits_{\mathbb{R}^p} [2q(x) - 1] \left(1_{x^\intercal \hat{\beta}_n < 0} - \mathbb{E}1_{x^\intercal \hat{\beta}_n < 0}\right) dP_X(x)\right)^2, \tag{2.2}$$

where $\mathbb{E}$ denotes the expectation over *iid* training sets of size n drawn from P. The form of $Var\left(\tau(\hat{c})\right)$ reveals that there are two scenarios in which $\tau(\hat{c})$ is highly variable. The first occurs when $x^\intercal \hat{\beta}_n$ is likely to be small relative to $Var(x^\intercal \hat{\beta}_n)$ over a *large* range of x where $q(x) \neq 1/2$. Notice that this might occur when the classifier does well but is subject to overfitting. The second scenario occurs when $x^\intercal \hat{\beta}_n$ is likely to be small relative to $Var(x^\intercal \hat{\beta}_n)$ over a *small* range of x where $q(x)$ is far from 1/2. In this scenario there may be little overfitting but the classifier may be far from the Bayes rule and hence of poor quality. Note that poor classifier performance and overfitting are hallmarks of small samples. In either case, $\tau(\hat{c})$ need not concentrate around $\mathbb{E}\tau(\hat{c})$.

In order to provide good intuition for the small sample case, we require an asymptotic framework wherein the test error $\tau(\hat{c})$ does not concentrate about $\mathbb{E}\tau(\hat{c})$, even in large samples. One way of achieving this is to permit $P(X^\intercal \beta^* = 0)$ to be positive where $\beta^* \triangleq \arg\min_{\beta \in \mathbb{R}^p} PL(X, Y, \beta)$. This ensures that for all $x \in \mathbb{R}^p$ that satisfy $x^\intercal \beta^* = 0$, the indicator function $1_{x^\intercal \hat{\beta}_n < 0} = 1_{x^\intercal \sqrt{n}(\hat{\beta}_n - \beta^*) < 0}$ never settles down to a constant but rather converges to a non-degenerate distribution. Furthermore, if for a non-null subset of these x's we have $q(x) \neq 1/2$, then $Var\left(\tau(\hat{c})\right)$ does not converge to zero. Hereafter we refer to this as the *non-regular framework*. This language is consistent with that of Bickel et al. (2001). However, unlike the usual notion of non-regularity the limiting distribution of $\sqrt{n}(\hat{\tau}(\hat{c}) - \tau(\hat{c}))$ depends not only on the value of β^* but also the marginal distribution of X.

To see why it is useful to consider approximations that are valid even in the non-regular asymptotic framework we consider simulated data, which we call the quadratic example. Here the generative model satisfies $P(X^\intercal\beta^* = 0) = 0$. Data are generated according to the following mechanism

$$\begin{aligned} X_1, X_2 &\sim_{iid} Unif[0,5] \\ \epsilon &\sim N(0, 1/4) \\ Y &= sign(X_2 - (4/25)X_1^2 - 1 + \epsilon). \end{aligned}$$

The working classifier is given by $\hat{c}(x) = sign(\hat{\beta}_{n0} + \hat{\beta}_{n1}x_1 + \hat{\beta}_{n2}x_2)$ where $\hat{\beta}_n$ is constructed using squared error loss $L(X, Y, \beta) \triangleq (1 - YX^\intercal\beta)^2$. In this example $\beta^* \approx (-.225, -317, .439)$ so that the continuity of X_1 and X_2 ensures that the regularity condition $P(X^\intercal\beta^* = 0) = 0$ is satisfied. Consider two seemingly reasonable, and commonly employed methods for constructing a confidence set. The first is the centered percentile bootstrap (CPB). The CPB confidence set is formed by bootstrapping the centered and scaled in-sample error $\sqrt{n}(\mathbb{P}_n - P)1_{YX^\intercal\hat{\beta}_n<0}$. Note that $\sqrt{n}(\mathbb{P}_n - P)1_{YX^\intercal\hat{\beta}_n<0} = \sqrt{n}(\hat{\tau}(\hat{c}) - \tau(\hat{c}))$ where $\hat{\tau}(\hat{c}) \triangleq \mathbb{P}_n 1_{YX^\intercal\hat{\beta}_n<0}$ is the in-sample error. More specifically, let $\hat{u}$ and $\hat{l}$ be the $1 - \gamma/2$ and $\gamma/2$ percentiles of

$$\sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)1_{YX^\intercal\hat{\beta}_n^{(b)}<0}, \tag{2.3}$$

where $\hat{\mathbb{P}}_n^{(b)} \triangleq n^{-1}\sum_{i=1}^n M_{ni}\delta_{(x_i,y_i)}$ is the bootstrap empirical measure with weights $(M_{n1}, M_{n2}, \ldots, M_{nn}) \sim Multinomial(n, \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$ and $\hat{\beta}_n^{(b)} \triangleq \arg\min_{\beta\in\mathbb{R}^p} \hat{\mathbb{P}}_n^{(b)} L(X, Y, \beta)$.

Then the $1-\gamma$ CPB interval is given by $[\hat{\tau}(\hat{c}) - \hat{u}/\sqrt{n}, \hat{\tau}(\hat{c}) - \hat{l}/\sqrt{n}]$. The second approach is based on the asymptotic approximation

$$\sqrt{n}(\mathbb{P}_n - P)1_{YX^\intercal\hat{\beta}_n<0} \approx N\left(0, (1 - P1_{YX^\intercal\beta^*<0})P1_{YX\beta^*<0}\right). \tag{2.4}$$

Thus the normal approximation confidence set is given by $\hat{\tau}(\hat{c}) \pm z_{1-\gamma}\sqrt{\frac{\hat{\tau}(\hat{c})(1-\hat{\tau}(\hat{c}))}{n}}$ (see the binomial approximation in Chung and Han 2009). If $P(X^\intercal\beta^* = 0) = 0$ then both methods can be shown to be consistent.

The left hand side of Figure 1 shows the estimated coverage using 1000 Monte Carlo iterations of the CPB with 1000 bootstrap resamples, and the normal approximation. Both methods severely undercover in small samples. This is especially troubling since (i) the problem is low-dimensional, (ii) the linear classifier is of relatively high quality, (for example if $n = 30$ the expected test error $\mathbb{E}\tau(\hat{c}) \approx .11$) and (iii) the regularity condition $P(X^\intercal\beta^* = 0) = 0$ is satisfied. Why do these methods fail? Neither method correctly captures the additional variation in the test error across training samples due to the non-smoothness of the test error. Since the generative model satisfies the condition $P(X^\intercal\beta^* = 0) = 0$, the variation across training sets eventually becomes negligible and the methods deliver the desired coverage for n large.

To illustrate the effect of non-smoothness on the coverage consider the problem of finding a confidence interval for the functional $\tau_{smoothed}(\hat{c}) \triangleq P(1 + exp(aY\hat{c}(X)))^{-1}$, where a is a positive free parameter. Notice that the size of a varies inversely with the smoothness of $\tau_{smooth}(\hat{c})$. A value of $a > 0$ gives the expectation of a sigmoid function and a value of $a = \infty$ corresponds to $\tau(\hat{c})$. Coverage for $a = 0.1, 1.0$, and 10 are given in the right hand side of Figure 1. Notice that coverage increases with the smoothness of the target $\tau_{smoothed}(\hat{c})$. The dramatic difference in coverage between

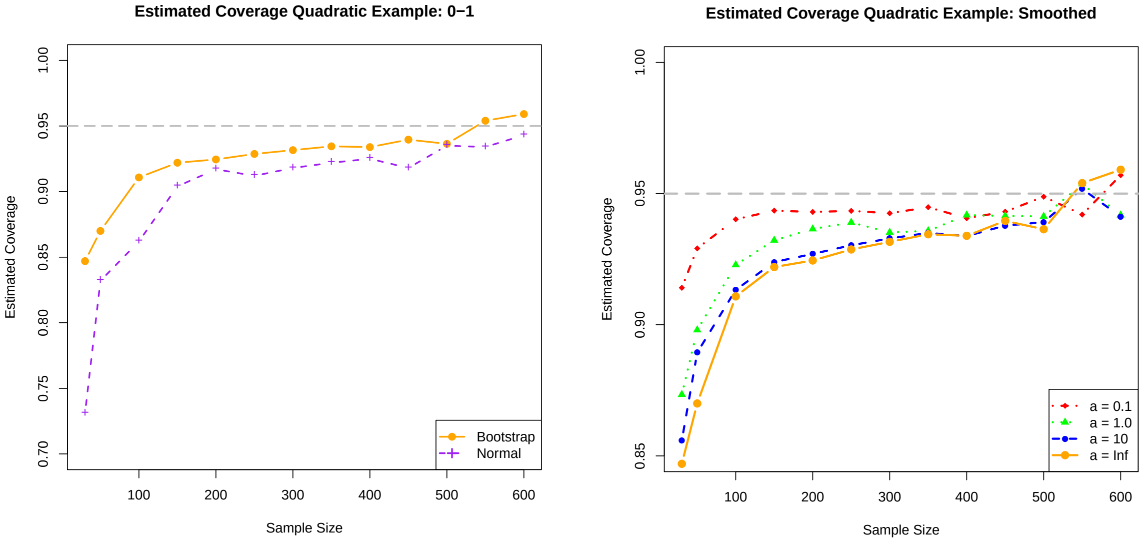

Figure 2.1: Left: Coverage of centered percentile bootstrap and normal approximations for constructing confidence sets for $\tau(\hat{c})$. Right: Coverage of centered percentile bootstrap with smoothed target $\tau_{smoothed}(\hat{c}) \triangleq P(1+exp(aY\hat{c}(X)))^{-1}$ for varying values of a; a value of $a = \infty$ corresponds to $\tau(\hat{c})$. Results are based on 1000 Monte Carlo iterations, target coverage is .950. The performance of the ACI on this example can found in Section 5 under the example labeled "quad."

$a = .1$ and $a = \infty$ suggests that a large component of the anti-conservatism is indeed attributable to non-smoothness.

Operating in the regular framework there is no indication that these methods may not work well. In the non-regular framework, however, both of these methods are inconsistent. To see this in the case of the CPB, write

$$\begin{aligned}\sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)1_{YX^\intercal\hat{\beta}_n^{(b)}<0} &= \sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)1_{X^\intercal\beta^*=0}1_{YX^\intercal\left[\sqrt{n}(\hat{\beta}_n^{(b)}-\hat{\beta}_n)+\sqrt{n}(\hat{\beta}_n-\beta^*)\right]<0} \\ &\quad + \sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)1_{X^\intercal\beta^*\neq 0}1_{YX^\intercal\hat{\beta}_n^{(b)}<0}. \end{aligned} \tag{2.5}$$

The first term on the right hand side of (2.5) appears because we allow $P(X^\intercal\beta^* =$

0) > 0 in the non-regular framework; conditioned on the data the term $\sqrt{n}(\hat{\beta}_n - \beta^*)$ does not have a limit and consequently the CPB is inconsistent. A detailed proof is omitted (see for example Shao 1994). The inconsistency of the normal approximation can be seen by examining the limiting distribution of $\sqrt{n}(\mathbb{P}_n - P)1_{YX^\intercal\hat{\beta}_n<0}$ in the non-regular framework. This limit is given in Theorem 3.1.

2.3 Adaptive confidence interval

In this section we introduce our method for constructing a confidence interval for the test error. This section is organized as follows. We begin by constructing adaptive confidence interval. Next, we establish the theoretical underpinnings of the method under fixed alternatives. Following this we provide a (heuristic) justification for our method using local alternatives. Finally, we discuss the choice of a tuning parameter required by the method.

2.3.1 Construction of the ACI

We propose an method of constructing a confidence interval that is consistent in the non-regular framework. We refer to this method as the Adaptive Confidence Interval (ACI) because, it is adaptive in two ways. First, unlike the CPB, the ACI provides asymptotically valid confidence intervals regardless of the true parameter values; intuitively the ACI achieves this by adapting to the amount of non-smoothness in the test error. Second, in settings (see Corollary 3.4) in which the CPB is consistent, the upper and lower limits of the ACI are adaptive in that these limits have the same distribution as the upper and lower limits of the CPB.

The ACI is based on bootstrapping an upper bound of the functional $\sqrt{n}(\mathbb{P}_n - P)1_{YX^\intercal\hat{\beta}_n<0}$. This upper bound is constructed by first partitioning the training data

$\mathcal{T}$ into two groups (i) points that are far from the boundary $x^\intercal\beta^* = 0$, and (ii) points that are too close to delineate from being on the boundary. The upper bound is constructed by taking the supremum over all possible classifications of the points that we cannot distinguish from lying on the boundary. More precisely, under the non-regular framework the scaled and centered test error can be decomposed as

$$\mathbb{G}_n 1_{YX^\intercal\hat{\beta}_n<0} = \mathbb{G}_n 1_{X^\intercal\beta^*=0} 1_{YX^\intercal\hat{\beta}_n<0} + \mathbb{G}_n 1_{X^\intercal\beta^*\neq 0} 1_{YX^\intercal\hat{\beta}_n<0}, \tag{2.6}$$

where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$. The first term on the right hand side of (2.6) corresponds to points on the decision boundary $x^\intercal\beta^* = 0$, and the second term corresponds to points that are not on this boundary. That is, the domain of X is partitioned into two-sets. We operationalize this partitioning using a series of hypothesis tests. For each $X = x$ we test $H_0 : x^\intercal\beta^* = 0$ against a two-sided alternative. Let Σ denote the asymptotic covariance of $\hat{\beta}_n$ (see below). Then the test rejects when the statistic $\frac{(x^\intercal\hat{\beta}_n)^2}{x^\intercal\Sigma x}$ is large. The bounds are obtained by computing the supremum (infemum) over all classifications of points for which the test fails to reject. In particular, an upper bound on $\mathbb{G}_n 1_{YX^\intercal\hat{\beta}_n<0}$ is given by

$$u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n) = \sup_{b\in\mathbb{R}^p} \mathbb{G}_n 1_{\frac{(X^\intercal\hat{\beta}_n)^2}{X^\intercal\Sigma X} \leq \frac{1}{a_n}} 1_{YX^\intercal b<0} + \mathbb{G}_n 1_{\frac{(X^\intercal\hat{\beta}_n)^2}{X^\intercal\Sigma X} > \frac{1}{a_n}} 1_{YX^\intercal\hat{\beta}_n<0}, \tag{2.7}$$

and an lower bound is given by

$$\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n) = \inf_{b\in\mathbb{R}^p} \mathbb{G}_n 1_{\frac{(X^\intercal\hat{\beta}_n)^2}{X^\intercal\Sigma X} \leq \frac{1}{a_n}} 1_{YX^\intercal b<0} + \mathbb{G}_n 1_{\frac{(X^\intercal\hat{\beta}_n)^2}{X^\intercal\Sigma X} > \frac{1}{a_n}} 1_{YX^\intercal\hat{\beta}_n<0}. \tag{2.8}$$

The choice of a_n, is discussed at the end of this Section. Put $b = \hat{\beta}_n$ to see that (2.7) and (2.8) are upper and lower bounds, respectively.

Suppose we want to construct a $1 - \delta\%$ confidence interval for the test error. We have that

$$\mathbb{P}_n 1_{YX^\intercal\hat{\beta}_n<0} - (1/\sqrt{n})u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n) \leq P1_{YX^\intercal\hat{\beta}_n<0} \leq \mathbb{P}_n 1_{YX^\intercal\hat{\beta}_n<0} - (1/\sqrt{n})\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n).$$

We approximate the distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$, $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ by bootstrap. The bootstrap is shown to be consistent later in this section. Denote the $1 - \delta/2$ percentile of the bootstrap distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ by $u_{1-\delta/2}$ and the $\delta/2$ percentile of the bootstrap distribution of $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ by $\ell_{\delta/2}$. The $1 - \delta\%$ ACI is given by

$$\mathbb{P}_n 1_{YX^\intercal\hat{\beta}_n<0} - (1/\sqrt{n})u_{1-\delta/2} \leq P1_{YX^\intercal\hat{\beta}_n<0} \leq \mathbb{P}_n 1_{YX^\intercal\hat{\beta}_n<0} - (1/\sqrt{n})\ell_{\delta/2}. \tag{2.9}$$

2.3.2 Properties of the ACI

In the remainder of the paper we verify that the ACI is asymptotically of the correct size even if the problem is non-regular (e.g. $P(X^\intercal\beta^* = 0) > 0$) and we evaluate the performance of the ACI in small samples. A method for efficiently approximating the ACI is given and shown to be almost identical to exact computation on a suite of examples. Most proofs are deferred to the appendix.

First we provide the asymptotic distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ and $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$. Throughout we make the following assumptions.

(A1) $L(X, Y, \beta)$ is convex with respect to β for each fixed $(x, y) \in \mathbb{R}^p \times \{-1, 1\}$.

(A2) $Q(\beta) \triangleq PL(X, Y, \beta)$ exists and is finite for all $\beta \in \mathbb{R}^p$.

(A3) $\beta^* \triangleq \arg\min_{\beta\in\mathbb{R}^p} Q(\beta)$ exists and is unique.

(A4) Let $g(X, Y, \beta)$ be a sub-gradient of $L(X, Y, \beta)$. Then $P||g(X, Y, \beta)||^2 < \infty$ for all β in a neighborhood of β^*.

(A5) $Q(\beta)$ is twice continuously differentiable at β^* and $H = \nabla^2 Q(\beta^*)$ is positive definite.

(A6) $\lim_{n\to\infty} a_n = \infty$ but $a_n = o(n)$.

These assumptions are quite mild and hold for most commonly used loss functions (e.g., exponential loss, squared error loss, hinge loss–if P has a smooth density at 1, logistic loss, etc.). Recall that a subgradient satisfies $L(x, y, \gamma) + (\beta - \gamma)^\intercal g(x, y, \gamma) \leq L(x, y, \beta)$ for all $(x, y) \in \mathbb{R}^p \times \{-1, 1\}$ and $\gamma, \beta \in \mathbb{R}^p$. All convex functions have a measurable subgradient. Let Ω be the covariance matrix of the sub-gradient of $L(x, y, \beta)$ at β^*. Under (A1)-(A5) Haberman (1989; see also Niemiro, 1992) proved that $\hat\beta_n$ converges with probability one to β^* and $\sqrt{n}(\hat\beta_n - \beta^*)$ converges in distribution to $z_\infty =_{\mathcal{L}} N(0, H^{-1}\Omega H^{-1})$.

Let $\mathbb{V}$ be a Brownian-Bridge indexed by $\mathbb{R}^p$ with the variance-covariance function

$$Cov(\mathbb{V}(\phi), \mathbb{V}(\gamma)) = P\left[1_{X^\intercal\beta^*=0}1_{YX^\intercal\phi<0} - P1_{X^\intercal\beta^*=0}1_{YX^\intercal\phi<0}\right] \times \left[1_{X^\intercal\beta^*=0}1_{YX^\intercal\gamma<0} - P1_{X^\intercal\beta^*=0}1_{YX^\intercal\gamma<0}\right]. \quad (2.10)$$

Furthermore, let $\mathbb{B}(\beta^*)$ denote a mean zero normal random variable with variance $P(1_{X^\intercal\beta^*\neq0}1_{YX^\intercal\beta^*<0} - P1_{X^\intercal\beta^*\neq0}1_{YX^\intercal\beta^*<0})^2$.

Theorem 2.3.1. *Let $\mathbb{V}$, $\mathbb{B}(\beta^*)$ and z_∞ be as above. Assume (A1)-(A6). Then*

1. $\mathbb{G}_n 1_{YX^\intercal\hat\beta_n<0} \rightsquigarrow \mathbb{V}(z_\infty) + \mathbb{B}(\beta^*)$,
2. $u(\mathbb{G}_n, \hat\beta_n, \Sigma, a_n) \rightsquigarrow \sup_{u\in\mathbb{R}^p} \mathbb{V}(u) + \mathbb{B}(\beta^*)$ *and* $\ell(\mathbb{G}_n, \hat\beta_n, \Sigma, a_n) \rightsquigarrow \inf_{u\in\mathbb{R}^p} \mathbb{V}(u) + \mathbb{B}(\beta^*)$.

Note that the limiting distributions of $u(\mathbb{G}_n, \hat\beta_n, \Sigma, a_n)$, $\ell(\mathbb{G}_n, \hat\beta_n, \Sigma, a_n)$ and $\mathbb{G}_n 1_{YX^\intercal\hat\beta_n<0}$ have the same regular component $\mathbb{B}(\beta^*)$; the three limits differ only in the non-regular component. Note also that the form of the covariance function of $\mathbb{V}$ given in (2.10) and the form of the limiting distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ (or $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$) shows that if the margin condition $P(X^\intercal\beta^* = 0) = 0$ holds, then $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n) \rightsquigarrow \mathbb{B}(\beta^*) =_{\mathcal{L}} lim_{n\to\infty}\mathbb{G}_n 1_{YX^\intercal\hat{\beta}_n<0}$ and similarly for $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$. That is, if the margin condition holds, the limiting distribution of the functional used to construct the ACI is the same as the limiting distribution of the functional $\mathbb{G}_n 1_{YX^\intercal\hat{\beta}_n<0}$. From a practical point of view this means that for problems where the regular framework is applicable, for example, if the sample size is large or points are well separated from the boundary, the ACI is asymptotically exact.

Another scenario in which the limiting distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$, $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ and $\mathbb{G}_n 1_{YX^\intercal\hat{\beta}_n<0}$ are the same is when the Bayes decision boundary is linear. In this case $q(x) = 1/2$ if $x^\intercal\beta^* = 0$ where $q(x) = P(Y = 1|X = x)$. (Here, we assume that the loss function is classification-calibrated (Bartlett 2005). All loss functions mentioned in this paper are classification-calibrated.) Then for any fixed $u \in \mathbb{R}^p$ we have

$$\begin{aligned}
P1_{X^\intercal\beta^*=0}1_{YX^\intercal u<0} &= \int\limits_{\{x\,:\,x^\intercal\beta^*=0\}} \left[q(x)1_{x^\intercal u<0} + (1-q(x))(1-1_{x^\intercal u<0})\right] dP_X(x) \\
&= \int\limits_{\{x\,:\,x^\intercal\beta^*=0\}} \left[2q(x)-1\right]1_{x^\intercal u<0}\, dP_X(x) + \frac{1}{2}P(X^\intercal\beta^*=0) \\
&= \frac{1}{2}P(X^\intercal\beta^*=0).
\end{aligned}$$

The form of the variance of $\mathbb{V}$ and the above series of equalities show that if the Bayes decision boundary is linear then $\mathbb{V}(u) =_{\mathcal{L}} N(0, \frac{1}{2}(1 - \frac{1}{2}P1_{X^\intercal\beta^*=0})P1_{X^\intercal\beta^*=0})$ for all $u \in \mathbb{R}^p$. Therefore, if the Bayes decision is linear

$$\begin{aligned} lim_{n\to\infty} u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n) &=_{\mathcal{L}} \sup_{u\in\mathbb{R}^p} \mathbb{V}(u) + \mathbb{B}(\beta^*) \\ &=_{\mathcal{L}} N\left(0, (1 - P1_{X^\intercal\beta^*<0})P1_{X^\intercal\beta^*<0}\right) + \mathbb{B}(\beta^*) \\ &=_{\mathcal{L}} \mathbb{V}(z_\infty) + \mathbb{B}(\beta^*) \\ &=_{\mathcal{L}} \lim_{n\to\infty} \sqrt{n}(\mathbb{P}_n - P)1_{YX^\intercal\hat{\beta}_n<0}, \end{aligned}$$

where the first and last equalities follow from Theorem 3.1, and the second and third equalities follow since $\mathbb{V}$ is constant across all indices. We have proved the following result.

Corollary 2.3.2. *Assuming (A1)-(A6) hold then if either (i) the Bayes decision boundary is $sign(X^\intercal\beta^*)$ or (ii) $P(X^\intercal\beta^* = 0) = 0$ then $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$, $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ and $\mathbb{G}_n 1_{YX^\intercal\hat{\beta}_n<0}$ have the same limiting distribution.*

The implication of the above theorem and corollary is that when either of the above conditions hold the ACI should provide the nominal coverage. When neither event holds then the ACI may be conservative. In simulations we shall see that the degree of conservatism is small.

The ACI in (2.9) utilizes a bootstrap approximation to the distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$, $\ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$. The next theorem concerns the consistency of the bootstrap distributions. Let $\hat{\Sigma}_n$ be a weakly consistent estimator of Σ (e.g. the plug-in estimator). Define $BL_1(\mathbb{R}^2)$ to be the space of bounded Lipschitz-1 functions on $\mathbb{R}^2$ and let $\mathbb{E}_M$ denote the expectation with respect to the bootstrap weights.

Theorem 2.3.3. *Assume (A1)-(A6). Then $\{u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n),\ \ell(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)\}$ and $\{u(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}_n, a_n),\ \ell(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}_n, a_n)\}$ converge to the same limiting distribution* *in probability. That is,*

$$\sup_{h\in BL_1(\mathbb{R}^2)}\Bigg|\mathbb{E}h\left(\{u(\mathbb{G}_n,\hat{\beta}_n,\Sigma,a_n),\ \ell(\mathbb{G}_n,\hat{\beta}_n,\Sigma,a_n)\}\right) \\ -\mathbb{E}_M h\left(\{u(\mathbb{G}_n^{(b)},\hat{\beta}_n^{(b)},\hat{\Sigma}_n,a_n),\ \ell(\mathbb{G}_n^{(b)},\hat{\beta}_n^{(b)},\hat{\Sigma}_n,a_n)\}\right)\Bigg|$$

converges in probability to zero.

Thus the ACI provides asymptotically valid confidence intervals. Moreover we have the following.

Corollary 2.3.4. *Assuming (A1)-(A6) hold then if either (i) the Bayes decision boundary is sign*$(X^\intercal\beta^*)$ *or (ii)* $P(X^\intercal\beta^*=0)=0$ *then* $u(\mathbb{G}_n^{(b)},\hat{\beta}_n^{(b)},\hat{\Sigma}_n,a_n)$, $\ell(\mathbb{G}_n^{(b)},\hat{\beta}_n^{(b)},\hat{\Sigma}_n,a_n)$ *and* $\mathbb{G}_n 1_{YX^\intercal\hat{\beta}_n<0}$ *converge to the same limiting distribution, in probability.*

Thus, the ACI is also adaptive in the sense that in settings where the centered percentile bootstrap *would* be consistent, $u(\mathbb{G}_n^{(b)},\hat{\beta}_n^{(b)},\hat{\Sigma},a_n)$, $\ell(\mathbb{G}_n^{(b)},\hat{\beta}_n^{(b)},\hat{\Sigma},a_n)$ and $\mathbb{G}_n 1_{YX^\intercal\hat{\beta}_n<0}$ have the same limiting distribution.

2.3.3 Local Alternatives

In Section 2 we motivated the use of a non-regular asymptotic framework in order to gain intuition for small samples. An alternative strategy for developing intuition for non-regular problems is to study the limiting behavior of $\sqrt{n}(\hat{\beta}_n-\beta^*)$ under local alternatives. This strategy has roots in Econometrics.

In econometrics, a common strategy to constructing procedures with good small sample properties in non-regular settings is to utilize alternatives local to the parameter values that cause the non-regularity (Andrews 2000; Cheng 2008; Xie 2009). To see this recall that in small samples a non-negligible proportion of the inputs x are in a $\sqrt{n}$-neighborhood of the decision boundary $x^\intercal\beta^*=0$ which causes the indicator function $1_{x^\intercal \hat{\beta}_n < 0}$ to become unstable. In the prior sections we assumed that there was a non-null probability that an input lies *exactly* on the boundary in order to retain the instability of the indicator function even in large samples. Another way to maintain this instability is by considering local alternatives.

The ACI can be seen as arising as an asymptotic approximation under local alternatives in the following way. In particular, suppose that a training set $\mathcal{T}_n = \{(X_{ni}, Y_{ni})\}_{i=1}^n$ is drawn *iid* from distribution P_n for which

$$\beta_n^* \triangleq \arg\min_{\beta \in \mathbb{R}^p} P_n L(X, Y, \beta) = \beta^* + \Gamma/\sqrt{n} \tag{2.11}$$

for some $\Gamma \in \mathbb{R}^p - \{0\}$. In addition, we assume that $P(X^\intercal \beta^* = 0) > 0$ (while $P_n(X^\intercal \beta_n^* = 0) > 0$ may or may not hold). A general tactic is to derive the limiting distribution of an estimator which will depend on the local parameter Γ and then take a supremum over this parameter to construct a confidence interval. As a first step in following this approach we might expect that

$$\begin{aligned} \mathbb{G}_n 1_{YX^\intercal \hat{\beta}_n < 0} = \mathbb{G}_n 1_{X^\intercal \beta^* = 0} 1_{YX^\intercal \left[\sqrt{n}(\hat{\beta}_n - \beta_n^*) + \Gamma\right] < 0} + \mathbb{G}_n 1_{X^\intercal \beta^* \neq 0} 1_{YX^\intercal \hat{\beta}_n < 0} \\ \rightsquigarrow \mathbb{V}(z_\infty + \Gamma) + \mathbb{B}(\beta^*) \end{aligned}$$

under P_n. Note that $\sup_\Gamma \mathbb{G}_n 1_{X^\intercal \beta^* = 0} 1_{YX^\intercal \left[\sqrt{n}(\hat{\beta}_n - \beta_n^*) + \Gamma\right] < 0}$ is equal to the first term on the right hand side of (2.7). Hence, $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ is the supremum over all local alternatives of the form given in (2.11). Also taking the supremum over $\Gamma \in \mathbb{R}^p - \{0\}$ we obtain

$$\sup_{\Gamma \in \mathbb{R}^p - \{0\}} \mathbb{V}(z_\infty + \Gamma) + \mathbb{B}(\beta^*) =_{\mathcal{L}} \sup_{u \in \mathbb{R}^p} \mathbb{V}(u) + \mathbb{B}(\beta^*),$$

which is the limiting distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ (see Theorem 3.1). Thus, the

ACI can be seen as arising as an asymptotic approximation under local alternatives. This result is formalized below.

Theorem 2.3.5. *Assume that $\mathcal{T}_n = \{(X_{ni}, Y_{ni})\}_{i=1}^n$ is drawn iid from distribution P_n for which:*

(B1) $\beta_n^* \triangleq \arg\min_{\beta \in \mathbb{R}^p} P_n L(X, Y, \beta) = \beta^* + \Gamma/\sqrt{n}$ *for some* $\Gamma \in \mathbb{R}^p - \{0\}$*,*

(B2) if $\mathcal{F}$ is any uniformly bounded Donsker class and $\mathbb{G}_n \rightsquigarrow \mathbb{L}$ in $l^\infty(\mathcal{F})$ under P, then $\mathbb{G}_n \rightsquigarrow \mathbb{L}$ in $l^\infty(\mathcal{F})$ under P_n,

(B3) $\sqrt{n}(\hat{\beta}_n - \beta_n^*) = -H^{-1}\mathbb{G}_n g(X, Y, \beta^*) + o_{P_n}(1)$*,*

where $\mathbb{G}_n \triangleq \sqrt{n}(\mathbb{P}_n - P_n)$. Assume (A1)-(A6). Then:

1. $\mathbb{G}_n 1_{YX^\intercal \hat{\beta}_n < 0} \rightsquigarrow \mathbb{V}(z_\infty + \Gamma) + \mathbb{B}(\beta^*)$

2. $\lim_{n\to\infty} u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n) =_{\mathcal{L}} \sup_{\eta \in \mathbb{R}^p} \mathbb{V}(z_\infty + \eta) + \mathbb{B}(\beta^*) = \sup_{u \in \mathbb{R}^p} \mathbb{V}(u) + \mathbb{B}(\beta^*)$

under P_n.

Thus the limiting distribution of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ is unchanged under local alternatives and hence might be expected to perform well in small samples. A similar result can be proved for $\ell(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}, a_n)$. This result is underscored by the empirical results in Section 5.

2.3.4 Choice of Tuning Parameter a_n

Use of the ACI requires the choice of the tuning parameter a_n. We use a simple heuristic for choosing the value of this parameter. The method described here performed well on all of the examples in Section 5. We begin with the presumption that undercoverage is a greater sin than conservatism. Recall that we can view the ACI as a two step procedure where at the first stage we test the null hypothesis $H_0 : x^{\intercal}\beta^* = 0$ against a two-sided alternative. The test of H_0 used in constructing the ACI rejects when $\frac{(X^{\intercal}\hat{\beta}_n)^2}{X^{\intercal}\Sigma X} > \frac{1}{a_n}$. The form of $u(\mathbb{G}_n, \hat{\beta}_n, \Sigma, a_n)$ in (2.7) shows that $\frac{1}{a_n}$ too small (e.g. large Type I error) results in too few points being deemed "near the boundary." Consequently the resulting interval may be too small since the supremum does not affect enough of the training points. Conversely, $\frac{1}{a_n}$ too large (e.g. large Type II error) puts too many points in the region on non-regularity, resulting in an interval that may be too wide because the supremum affects too many of the training points. Given our presumption, controlling Type I error is of primary importance. Let $\gamma \in (0, 1)$. Then let $\frac{1}{a_n} = \frac{1}{\sqrt{n}} \vee \frac{\chi^2_{1-\gamma}}{n}$ and we have for any $x \in \mathbb{R}^p - \{0\}$ and $x^{\intercal}\beta^* = 0$

$$P\left(\frac{(x^{\intercal}\hat{\beta}_n)^2}{x^{\intercal}\Sigma x} > \frac{1}{a_n} \middle| H_0 \right) = P\left(\left(\frac{\sqrt{n}(\hat{\beta}_n - \beta^*)^{\intercal} x}{\sqrt{x^{\intercal}\Sigma x}} \right)^2 > \frac{n}{a_n} \right) \lesssim \gamma.$$

Thus, the suggested a_n controls the Type I error to be no more than γ. Moreover, it is clear from the above display that the Type I error decreases to zero as n tends to infinity. In all of the experiments in this paper we choose, rather arbitrarily, to use $\gamma = .005$. Simulations results, given in Table 5 of the appendix, show that the performance (measured in terms of width and coverage) of the ACI appears to be insensitive to choices of γ in the range .001 to .01 for a sample size of around 30. For larger sample sizes, the choice of γ is unimportant since $\sqrt{n} > \chi^2_{1-\gamma}$ except for extremely small values of γ.

2.4 Computation

To implement the ACI we need to calculate, for each bootstrap sample, the supremum and infimum in $u(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}_n, a_n)$, and $l(\mathbb{G}_n^{(b)}, \hat{\beta}_n^{(b)}, \hat{\Sigma}_n, a_n)$ respectively. The required optimization, as stated, is a Mixed Integer Program (MIP) because of the discrete nature of the indicator function. In this section, we develop a convex relaxation that can be solved in polynomial time. The details for the infimum are provided below; a similar approach is used to find the supremum by writing $1_{z<0} = 1 - 1_{z\geq 0}$ and using the relationship: $\sup_z g(z) = -inf_z - g(z)$. Let $(m_{n1}, m_{n2}, \ldots, m_{nn})$ be a realization of the bootstrap weights $(M_{n1}, M_{n2}, \ldots, M_{nn}) \sim Multinomial(n, \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$. For each such realization, construction of the infemum in the ACI requires computing

$$\inf_{u\in\mathbb{R}^p} \sum_{i\in N_n^{(b)}} (m_{ni} - 1)1_{y_i x_i^\intercal u<0}, \tag{2.12}$$

where $N_n^{(b)} = \{i : \frac{(x_i^\intercal \hat{\beta}_n^{(b)})^2}{x_i^\intercal \hat{\Sigma}_n x_i} \leq \frac{1}{a_n}\}$. In this form, the optimization is clearly seen to be an MIP. Reliably solving an MIP requires the use specialized software (we use CPLEX) and quickly becomes computationally burdensome as the size of the problem grows. The following convex relaxation of (2.12) is (i) computationally efficient requiring roughly the same amount of computation as fitting a linear SVM and (ii) can be solved without specialized software (e.g. R or matlab).

As the initial step write

$$\sum_{i\in N_n^{(b)}} (m_{ni} - 1)1_{y_i x_i^\intercal u<0} = \sum_{i\in N_n^{(b)}} m_{ni} 1_{y_i x_i^\intercal u<0} + \sum_{i\in N_n^{(b)}} (-1_{y_i x_i^\intercal u<0}).$$

Then replace the indicator function $1_{y_i x_i^\intercal u<0}$ with convex surrogate and upper bound $(1 - y_i x_i^\intercal u)_+$ where $(z)_+$ denotes the positive part of z. Similarly, replace the function $-1_{y_i x^\intercal u<0}$ with convex surrogate and upper bound $(1 + y_i x_i^\intercal u)_+ - 1$. The indicator functions and their respective surrogates are shown in Figure 2. The relaxed

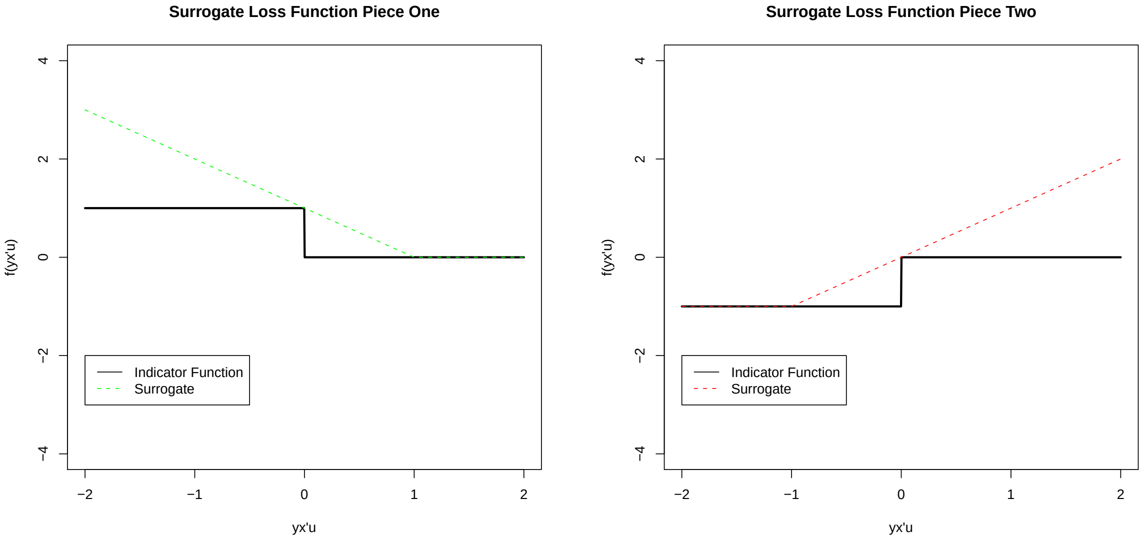

Figure 2.2: Relaxation of the indicator functions. Left panel: indicator function $1_{yx^\intercal u<0}$ replaced with convex surrogate $(1-yx^\intercal u)_+$. Right panel: indicator function $-1_{yx^\intercal u<0}$ replaced with convex surrogate $(1+yx^\intercal u)_+ - 1$.

optimization problem is then

$$\inf_{u\in\mathbb{R}^p} \sum_{i\in N_n^{(b)}} [m_{ni}(1-y_i x_i^\intercal u)_+ + (1+y_i x_i^\intercal u)_+] \tag{2.13}$$

where the -1 in the relaxation of $-1_{y_i x_i^\intercal u<0}$ has been omitted since it does not depend on u. The optimization problem in (2.13) can be cast as a linear program and hence solved in polynomial time. See the next section for an empirical comparison of the relaxed and MIP solutions to (2.12).

2.5 Empirical study

In this section we compare solution quality between the relaxed and MIP solutions to (2.12); as will be seen the relaxed solution to (2.12) can be computed much more quickly while little is lost in terms of solution quality. Next using the relaxed solution to (2.12) the empirical performance of the ACI is compared with two recent methods proposed in the literature. Ten data sets are used in these comparisons; three are simulated and the remaining seven data sets are taken from the UCI machine learning repository (www.ics.uci.edu/∼mlearn/MLRepository.html) and thus the true generative model is unknown. In this case, the empirical distribution function of the data set is treated as the generative model. A summary of the data sets are given in Table 2.2.

To assess the difference in solution quality between the relaxed and MIP solutions to (2.12) we perform the following procedure for each of the 10 examples listed in Table 2.2. We generate 1000 training sets of size $n = 30$, and for each training set we compute 1000 bootstrap resamples. For each resample we compute (2.12) exactly using the MIP and approximately using the convex relaxation described above. Here we illustrate the results when the loss function used to construct $\hat{\beta}_n$ and $\hat{\beta}_n^{(b)}$ is chosen to be $L(X, Y, \beta) = (1 - YX^{\intercal}\beta)^2$. Let $\theta_{MIP}^{(t)(b)}$ and $\theta_{REL}^{(t)(b)}$ denote the MIP and relaxed solution to (2.12) for the b^{th} bootstrap resample of the t^{th} training set. Table 2.1 reports the 50, 75, 95, and 99 percentiles of $\frac{1}{n}\left(\theta_{MIP}^{(t)(b)} - \theta_{REL}^{(t)(b)}\right)$ for each example. Notice that for each example we considered, the relaxed and MIP solutions agree exactly on more than half of the resampled pairs. Moreover, on more than 95 percent of the resampled pairs, we observe that $\frac{1}{n}\left(\theta_{MIP}^{(t)(b)} - \theta_{REL}^{(t)(b)}\right) \leq \frac{1}{n}$, implying that the two solutions differed by at most the activation of a single indicator function. Table 2.1 also reports the estimated coverage of confidence sets constructed using the MIP and relaxed formulations. For each of the 10 data sets, estimated coverage using the two methods is not significantly different. The final bit of information in Table 2.1 regards computation time. The last two columns report the average time in seconds that it takes to construct a single confidence interval using the MIP and relaxed formulations. Computations were performed using a 3.06 GHz intel processor with 4

GB 1067 MHz DDR3. It is clear that even in the $n = 30$ case significant computational gain can be made by using the relaxed formulation. However, this gain becomes more pronounced as sample size increases. Figure 3 compares the computation time for the ThreePt data set (this data set is decribed in Laber and Murphy 2009) as a function of sample size using squared error loss. As claimed, the computation time for the relaxed construction scales much more efficiently than the MIP formulation. In the examples presented in the next section we use the convex relaxation to compute the confidence interval.

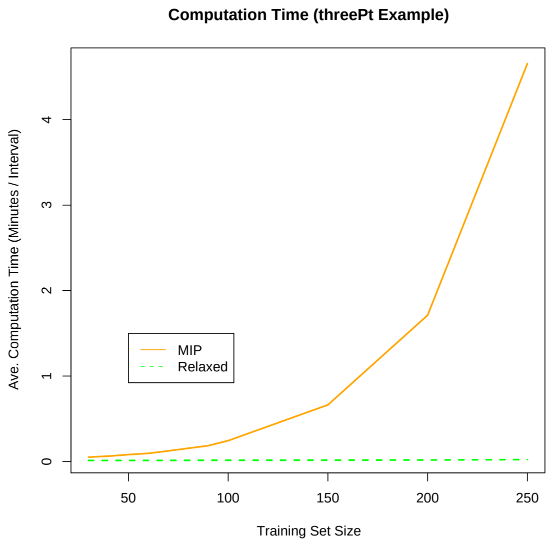

Figure 2.3: Computation time for MIP and relaxed construction of ACI using the ThreePt data set and squared error loss.

2.5.1 Competing methods

As competitors we consider a repeated-split normal approximation suggested by (Yang 2006) and the recently proposed Bootstrap Case Cross-Validated Percentile with Bias Reduction (BCCVP-BR) method of (Jiang 2008). These methods represent

	Coverage		Difference in width				Computation time	
Data Set	Relaxed	MIP	$p_{.99}$	$p_{.95}$	$p_{.75}$	$p_{.5}$	Relaxed	MIP
ThreePt	.948	.948	.0334	0.00	0.00	0.00	.734	3.11
Magic	.944	.945	.0334	.0334	0.00	0.00	1.24	1.94
Mam.	.957	.958	.0334	0.00	0.00	0.00	.904	1.88
Ion.	.954	.954	.0334	0.00	0.00	0.00	1.33	3.06
Donut.	.967	.968	.0667	.0334	.0334	0.00	.917	2.94
Bal.	.969	.969	0.00	0.00	0.00	0.00	.977	1.69
Liver	.956	.956	.0333	.0333	0.00	0.00	1.61	2.50
Spam	.984	.987	.0333	.0333	0.00	0.00	1.54	3.01
Quad	.959	.962	.0333	0.00	0.00	0.00	.983	1.37
Heart	.960	.961	.0333	0.00	0.00	0.00	1.06	3.27

Table 2.1: Comparison of MIP and relaxed versions of the ACI. For each data set the table was constructed using 1000 training sets each with 1000 bootstrap iterations for a total of 1,000,000 computations of the optimization problem given in (2.12).

the best we could find in terms of consistent coverage. Both methods substantially outperform standard approaches like the bootstrap and normal approximation which are discussed in Section 2. To provide a baseline for comparison, the performance of the Centered Percentile Bootstrap (CPB) is included in the appendix.

Briefly, Yang's method repeatedly partitions the training data $\mathcal{T}$ into two equal halves $\mathcal{T}^L$ and $\mathcal{T}^V$. A classifier is trained on $\mathcal{T}^L$ and then evaluated on $\mathcal{T}^V$. The mean and variance of the number of misclassified points in $\mathcal{T}^V$ is recorded. This mean and variance are then aggregated and used in a normal approximation. Jiang's method can be roughly described as leave one out cross validation with bootstrap resamples. However, since a bootstrap resample can have multiple copies of a single training example, leave one out cross-validation will no longer have disjoint training and testing sets. Instead, for each unique training example (x_i, y_i) the bootstrap resample is partitioned into two sets, one with all copies of (x_i, y_i) call this $\mathcal{V}$, and the second contains the remainder of the resample call this $\mathcal{L}$. The classifier is trained on $\mathcal{L}$ and evaluated on $\mathcal{V}$. The average error over all sets $\mathcal{V}$ is recorded within each bootstrap resample and the percentiles form the endpoints of a confidence interval. As a final step Jiang provides a bias correction. A full description of these methods can be found in the referenced works. While these methods are intuitive, they lack theoretical justification. Yang's method was developed for use with a hold-out set; when such a hold-out set does not exist, the method is inconsistent. Jiang offers no justification other than intuition.

2.5.2 Results

We examine the performance of the ACI and competing methods using the following three metrics (i) coverage (ii) interval width and (iii) computational expense. These metrics are recorded using ten data sets, three sample sizes, and three loss functions. Three of the examples use simulated datasets and hence the test error can be computed exactly. The remaining seven data sets are taken from the UCI machine learning repository
(www.ics.uci.edu/∼mlearn/MLRepository.html) and thus the true generative model is unknown. In this case, the empirical distribution function of the data set is treated as the generative model. Results using squared error loss are listed here while the results using binomial deviance and ridged hinge loss (support vector machines) are given in the appendix. A summary of the data sets are given in Table 2.2.

Coverage results for squared error loss are given in Table 2.3. The adaptive confidence interval is the only method to attain at least nominal coverage on all ten test sets. Yang's method is either extremely conservative or anti-conservative. Jiang's interval attains the nominal coverage on eight of ten data sets in the $n = 30$ case and nine of ten data sets for larger sample sizes. Table 2.4 shows the width of the constructed confidence intervals. When $n = 30$ the ACI is smallest in width for eight of the ten data sets. For larger sample sizes Jiang's method and the ACI display comparable widths; Yang's method is always the widest. Another important factor is computation time. Table 2.5 shows the average amount of time required in seconds to construct a single confidence interval. All methods used 1000 resamples. That is, 1000 bootstrap resamples for the ACI and Jiang's method, and 1000 repeated splits for Yang's method. Table 2.5 shows that Yang's method is the most computationally efficient. However, it is also clear that Jiang's method is significantly slower than the ACI for moderate sample sizes. For the Magic data set Jiang's method takes more than 30 times longer than the ACI. It is most important, however, to notice the trend in computation time across sample sizes. Computation time for Yang's method and the ACI grow slowly with sample size while the computational cost of Jiang's method increases much more quickly. The reason for this is that Jiang's method performs leave-one-out cross validation for each bootstrap resample thus increasing the computation time by a factor of n. Results for ridged hinge loss and binomial deviance loss are similar and can be found in the technical report (Laber and Murphy, 2010).

Name	Features	Source	$\mathbb{E}\tau(\hat{c})$ (SE)	$\mathbb{E}\tau(\hat{c})$ (BD)	$\mathbb{E}\tau(\hat{c})$ (SVM)
ThreePt	2	Simulated	.500	.500	.500
Quad	3	Simulated	.0997	.109	.101
Donut	3	Simulated	.235	.249	.232
Magic	11	UCI	.264	.231	.252
Mam.	6	UCI	.192	.190	.203
Ion.	9	UCI	.151	.147	.149
Bal.	5	UCI	.054	.050	.061
Liver	7	UCI	.342	.342	.334
Spam	10	UCI	.190	.183	.181
Heart	9	UCI	.167	.173	.174

Table 2.2: Test data sets used to evaluate confidence interval performance. The last three columns record the average test error for a linear classifier trained using a training set of size $n = 100$ and loss function: squared error loss (SE), binomial deviance (BD), and ridged hinge loss (SVM).

Sample Size	$n = 30$			$n = 100$			$n = 250$		
Data Set / Method	ACI	Yang	Jiang	ACI	Yang	Jiang	ACI	Yang	Jiang
ThreePt	.948	.930*	.863*	.937	.537*	.925*	.935	.387*	.930*
Magic	.944	.996*	.979*	.973*	.991*	.969*	.962	.996*	.974*
Mam.	.957	.989*	.966	.937	.996*	.964	.960	.995*	.968
Ion.	.941	.996*	.972*	.961	.992*	.964	.952	.996*	.949
Donut	.965	.967	.908*	.970*	.866*	.974*	.974*	.895*	.988*
Bal.	.976*	.989*	.966	.962	.995*	.969*	.946	.991*	.963
Liver	.956	.997*	.970*	.963	.992*	.966	.971*	.996*	.984*
Spam	.984*	.998*	.975*	.967	.996*	.967	.979*	.996*	.958
Quad	.959	.983*	.945	.957	.989*	.938	.965	.999*	.940
Heart	.960	.995*	.976*	.949	.991*	.979*	.971*	.989*	.974*

Table 2.3: Coverage comparison between ACI, Yang's CV and Jiang's $BCCVP - BR$ for squared error loss, target coverage is .950. Coverage is starred if observed coverage is significantly different from .950 at .01 level.

Sample Size	$n = 30$			$n = 100$			$n = 250$		
Data Set / Method	ACI	Yang	Jiang	ACI	Yang	Jiang	ACI	Yang	Jiang
ThreePt	.385*			.198*			.193*		
Magic	.498*	.528	.501	.238	.257	.214*	.125	.157	.122*
Mam.	.374*	.456	.383	.191	.226	.178*	.112	.140	.105*
Ion.	.313*	.466	.388	.175	.213	.172*	.103	.127	.100*
Donut	.424*	.483		.217*	.258		.123*		.201
Bal.	.217*	.350	.232	.101*	.138	.103	.0623	.0772	.0620*
Liver	.534	.527	.500*	.262	.274	.241*	.152	.172	.143*
Spam	.428	.496	.418*	.219	.229	.184*	.125	.140	.108*
Quad	.246*	.360	.267	.142*	.171	.144	.0811*	.104	.0885
Heart	.367*	.476	.404	.184*	.219	.184*	.106*	.132	.110

Table 2.4: Comparison of interval width between ACI, Yang's CV and Jiang's $BCCVP - BR$ for squared error loss. Smallest observed width is starred. Examples where at least the nominal coverage was not attained are omitted.

2.6 Discussion

Many statistical procedures in use today are justified by a combination of asymptotic approximations and high quality simulation performance. As exemplified here,

Sample Size	$n = 30$			$n = 100$			$n = 250$		
Data Set / Method	ACI	Yang	Jiang	ACI	Yang	Jiang	ACI	Yang	Jiang
ThreePt	.734			.762			1.37		
Magic	1.24	.0392	1.59	1.40	.0834	11.1	1.90	0.178	60.66
Mam.	1.37	.0185	.697	6.03	.0383	5.52	12.8	.0800	26.3
Ion.	2.13	.0331	1.32	6.42	.0702	10.0	16.7	.147	52.62
Donut	2.00	.00930		4.33		2.16	11.6		10.84
Bal.	.977	.0160	.575	1.05	.0315	3.50	1.23	.0660	20.9
Liver	1.16	.0222	.859	1.44	.0461	6.25	1.78	.0978	33.7
Spam	1.38	.0348	1.37	1.53	.0744	10.5	1.72	.159	57.9
Quad	.983	.00918	.125	1.11	.0191	1.43	1.24	.0398	6.96
Heart	1.06	.0317	1.25	1.15	.0660	8.00	1.42	.139	23.6

Table 2.5: Comparison of computation time (in seconds) between ACI, Yang's CV and Jiang's $BCCVP-BR$ for squared error loss. Examples where at least the nominal coverage was not attained are omitted.

the choice of asymptotic framework may be crucial in obtaining reliably good performance in small samples. In this paper a non-regular asymptotic framework in which the limiting distribution of the test error changes abruptly with changes in the true, underlying data generating distribution is used to develop a confidence interval. In particular, asymptotic non-regularity occurs due to the non-smooth test error in connection with particular combinations of β^* values and the X distribution. It is common practice to "eliminate" this asymptotic non-regularity by assuming that these problematic combinations of β^* values and the X distribution cannot occur. However, small samples are unable to precisely discriminate between settings that are *close to* the problematic β^* values/X distribution from settings in which the β^* values/X distribution are *exactly* problematic. As a result, asymptotic approximations that depend on assuming away these problematic settings can be of poor quality; this is the case here.

The validity of proposed adaptive confidence interval presented here does not depend on assuming away problematic scenarios; instead the ACI detects and then accommodates settings that are sufficiently close to the problematic β^* values/X distribution. In this sense the ACI adapts to the non-smoothness in the test error. Specifically, in settings in which standard asymptotic procedures fail, the ACI provides asymptotically valid, albeit conservative, confidence intervals. Moreover, the ACI delivers exact coverage if either (i) the model space is correct or (ii) a margin condition holds. Practically, this means that in a setting where standard asymptotic procedures (e.g. the bootstrap) are applicable, the ACI is asymptotically equivalent to these methods. Experimental performance of the ACI is also quite promising. On a suite of 10 examples, three loss functions and three classification algorithms, the ACI delivered nominal coverage. In addition, the ACI generally had a smaller length than competing methods. The ACI can be computed efficiently with algorithms scaling polynomially in dimension and sample size.

Two important extensions of the ACI are: first, to extend the ACI to construct valid confidence intervals for the difference in test error between two linear classifiers and, second, to extend these ideas to the setting in which the number of features is comparable or larger than the sample size. The former extension is straightforward and can be achieved by enlarging the set over which the supremum is taken in (2.7) to include the points on the classification boundaries of both classifiers. The latter is more difficult. In the estimation of classifiers in the $p >> n$ setting, it is important to avoid overfitting. A typical approach to reduce the amount of overfitting is regularization which effectively reduces the space of available classifiers to choose from. Similarly, the supremum in (2.7) must be taken over a restricted set of classifiers to avoid being unnecessarily wide. Extending the theory and computation to this setting is left to another paper.

2.7 Appendix I: Additional experimental results

Sample Size	$n = 30$			$n = 100$			$n = 250$		
Data Set / Method	ACI	Yang	Jiang	ACI	Yang	Jiang	ACI	Yang	Jiang
ThreePt	.976	.893*	.914*	.961	.552*	.945	.961	.387*	.930*
Magic	.955	.999*	.983*	.977*	.991*	.969*	.972*	.997*	.974*
Mam.	.957	.989*	.966	.962	.996*	.964	.960	.995*	.968
Ion.	.947	.995*	.985*	.948	.996*	.970*	.970	.990*	.970*
Donut	.968	.966	.908*	.969	.851*	.971*	.973*	.898*	.966
Bal.	.979*	.996*	.972*	.988*	.999*	.982*	.976*	.996*	.976*
Liver	.946	.995*	.972*	.973*	.995*	.974*	.977*	.993*	.984*
Spam	.985*	.999*	.981*	.983*	.996*	.967	.983*	.990*	.973*
Quad	.978*	.997*	.945	.957	.989*	.938	.965	.999*	.940
Heart	.960	.995*	.976*	.949	.991*	.979*	.971*	.989*	.974*

Table 2.6: Coverage comparison between ACI, Yang's CV and Jiang's $BCCVP{-}BR$ for binomial deviance loss, target coverage is .950.

Sample Size	$n = 30$			$n = 100$			$n = 250$		
Data Set / Method	ACI	Yang	Jiang	ACI	Yang	Jiang	ACI	Yang	Jiang
ThreePt	.374*			.191*			.119*		
Magic	.466*	.526	.504	.240*	.257	.241	.125	.158	.122*
Mam.	.374*	.456	.383	.191	.226	.178*	.112	.140	.105*
Ion.	.305*	.459	.402	.184	.212	.176*	.0998*	.127	.104
Donut	.434*	.485		.222*		.311	.124*		.205
Bal.	.263	.351	.257*	.125*	.148	.133	.0723*	.0820	.0784
Liver	.530	.526	.520*	.259	.274	.246*	.151	.171	.143*
Spam	.454	.494	.423*	.222	.234	.185*	.125	.141	.110*
Quad	.310	.373	.267*	.142*	.171	.144	.0811*	.104	.0885
Heart	.367*	.476	.404	.184*	.219	.184*	.102*	.132	.105

Table 2.7: Comparison of interval width between ACI, Yang's CV and Jiang's $BCCVP - BR$ for binomial deviance loss. Entries with the smallest width are starred. Examples where at least the nominal coverage was not attained are omitted.

Sample Size	$n = 30$			$n = 100$			$n = 250$		
Data Set / Method	ACI	Yang	Jiang	ACI	Yang	Jiang	ACI	Yang	Jiang
ThreePt	.950	.916*	.886*	.951	.886*	.893	.961	.492*	.905*
Magic	.952	.988*	.955	.989*	.996*	.978*	.976*	.996*	.938
Mam.	.958	.980*	.965	.962	.992*	.971*	.939	.990*	.968*
Ion.	.944	.990*	.975*	.968*	.954	.960	.970*	.964	.978*
Donut	.945	.986*	.912*	.939	.990*	.989*	.971*	.972*	.959
Bal.	.974*	.956	.970*	.998*	.862*	.973*	.963	.768*	.942
Liver	.963	.998*	.962	.985*	.984*	.953	.977*	.976*	.983*
Spam	.985*	.992*	.962	.954	.992*	.961	.973*	.992*	.985*
Quad	.971*	.997*	.971*	.960	.989*	.987*	.964	.999*	.967*
Heart	.958	.995*	.986*	.938	.991*	.959	.962	.989*	.934*

Table 2.8: Coverage comparison between ACI, Yang's CV and Jiang's $BCCVP{-}BR$ for ridged (L_2 penalized) hinge loss, target coverage is .950.

Sample Size	$n = 30$			$n = 100$			$n = 250$		
Data Set / Method	ACI	Yang	Jiang	ACI	Yang	Jiang	ACI	Yang	Jiang
ThreePt	.246*			.145*			.129*		
Magic	.487*	.514	.504	.228*	.253	.254	.131	.154	.130*
Mam.	.373*	.468	.422	.191*	.233	.199	.080*	.141	.105
Ion.	.313*	.457	.345	.171*	.221	.191	.102*	.129	.117
Donut	.335*	.463		.203*	.239	.399	.129*	.148	.214
Bal.	.239*	.370	.271	.126		.121*	.0841		.0723*
Liver	.456*	.520	.526	.252*	.273	.302	.154*	.170	.182
Spam	.454	.471	.371*	.206	.227	.196*	.116	.131	.109*
Quad	.283	.399	.291*	.158	.189	.140*	.0871	.111	.0857*
Heart	.387*	.498	.398	.191*	.229	.218	.111*	.135	

Table 2.9: Comparison of interval width between ACI, Yang's CV and Jiang's $BCCVP - BR$ for ridged (L_2 penalized) hinge loss. Entries with the smallest width are starred. Examples where at least the nominal coverage was not attained are omitted.

	Coverage					Width				
Data Set / γ	.01	.0075	.005	.0025	.001	.01	.0075	.005	.0025	.001
ThreePt	.958	.950	.948	.950	.946	.380	.381	.385	.385	.398
Magic	.944	.952	.944	.971	.968	.484	.493	.498	.501	.519
Mam.	.939	.944	.957	.952	.944	.360	.364	.374	.381	.384
Ion.	.926	.936	.941	.956	.960	.319	.319	.313	.339	.347
Donut	.944	.969	.965	.956	.956	.427	.427	.424	.437	.441
Bal.	.958	.972	.976	.982	.974	.212	.210	.217	.221	.223
Liver	.967	.953	.956	.946	.966	.527	.526	.534	.538	.548
Spam	.986	.982	.984	.985	.991	.434	.433	.428	.449	.451
Quad	.950	.942	.959	.942	.958	.244	.241	.246	.248	.252
Heart	.943	.953	.960	.945	.963	.356	.366	.367	.371	.379

Table 2.10: Coverage and width of ACI using squared error loss and a training set size of $n = 30$. Coverage and width appears to be stable across a range of reasonable values for tuning parameter γ.

Sample Size	n=30			n=100			n=250		
Data Set / Loss	Squared	Log.	Hinge	Squared	Log.	Hinge	Squared	Log.	Hinge
ThreePt	.862*	.781*	.902*	.910*	.872*	.918*	.918*	.936	.930*
Magic	.736*	.720*	.716*	.880*	.888*	.891*	.915*	.909*	.928*
Mam.	.795*	.807*	.790*	.882*	.881*	.905*	.912*	.927*	.942
Ion.	.799*	.775*	.753*	.876*	.878*	.881*	.906*	.975*	.953
Donut	.880*	.903*	.851*	.929*	.921*	.870*	.926*	.926*	.909*
Bal.	.872*	.925*	.835*	.889*	.932*	.878*	.883*	.931*	.889*
Liver	.758*	.746*	.727*	.868*	.851*	.847*	.910*	.960	.917*
Spam	.814*	.809*	.749*	.883*	.878*	.882*	.932*	.938	.913*
Quad	.849*	.918*	.849*	.908*	.926*	.922*	.928*	.954	.932*
Heart	.755*	.782*	.704*	.873*	.874*	.846*	.911*	.937	.930*

Table 2.11: Coverage of Centered Percentile Bootstrap (CPB) using three loss functions (squared loss, logistic loss, and hinge loss) and three sample sizes ($n = 30$, $n = 100$, and $n = 250$). Target coverage is .950, coverage is starred if significantly different from target.

2.8 Appendix II: Proofs

In this section we provide sketches of the results in the main body of the paper. Longer versions of these proofs are available in a technical report (see Laber and Murphy 2009). Please see the original statements for a list of assumptions and the notation. Before proceeding, some initial remarks are warranted. We assume that an intercept term has been included in the model so that $P_X(X \equiv 0) = 0$. In addition, if $\mathbb{T}$ is Brownian Bridge indexed by a class of functions $\mathcal{E}$, then we denote by $\rho_{\mathbb{T}}$ the covariance semi-metric on $\mathcal{E}$ given by $\rho_{\mathbb{T}}(e, d) \triangleq P(\mathbb{T}(e) - \mathbb{T}(d))^2$.

Proof. [Corollary 3.2 part (1)] To begin write

$$\sqrt{n}(\mathbb{P}_n - P)1_{YX^\intercal\hat{\beta}_n<0} = \sqrt{n}(\mathbb{P}_n - P)1_{X^\intercal\beta^*=0}1_{YX^\intercal\hat{\beta}_n<0} + \sqrt{n}(\mathbb{P}_n - P)1_{X^\intercal\beta^*\neq0}1_{YX^\intercal\hat{\beta}_n<0}.$$

We begin by establishing the limiting distribution of the first term on the right hand side of the above display. The class of functions $\mathcal{F} \triangleq \{1_{X^\intercal\beta^*=0}1_{YX^\intercal\gamma<0} : \gamma \in \mathbb{R}^p\}$ is a P-measurable, uniformly bounded VC class (see Anthony and Bartlett 2002 Theorem 8.14) and hence Donsker. Thus, if we define $\hat{\mathbb{V}}_n(\gamma) \triangleq \sqrt{n}(\mathbb{P}_n - P)1_{X^\intercal\beta^*=0}1_{YX^\intercal\gamma<0}$ then $\hat{\mathbb{V}}_n \rightsquigarrow \mathbb{V}$ in $l^\infty(\mathbb{R}^p)$. Moreover, under assumptions (A1)-(A5) $\sqrt{n}(\hat{\beta}_n - \beta^*) \rightsquigarrow N(0, \Sigma)$. Recall that $\Sigma \triangleq H^{-1}\Omega H^{-1}$ where $H \triangleq \nabla^2 Q(\beta^*)$ and $\Omega \triangleq Pg(X, Y, \beta^*)g(X, Y, \beta^*)^\intercal$. Moreover, since $\sqrt{n}(\hat{\beta}_n - \beta^*) = -H^{-1}\sqrt{n}(\mathbb{P}_n - P)g(X, Y, \beta^*) + o_P(1)$ (see Niemiro 1992; Haberman 1989) and for any fixed $\alpha \in \mathbb{R}^p$ the class $\mathcal{F} + \alpha^\intercal g(X, Y, \beta^*)$ is Donsker, finite dimensional convergence of $(\hat{\mathbb{V}}_n, \sqrt{n}(\hat{\beta}_n - \beta^*))$ is established by means of the Cramer-Wold device. Moreover, asymptotic tightness and measurability follow from Theorem 7.14 in Kosorok (2008). Thus, joint convergence of $(\hat{\mathbb{V}}_n, \sqrt{n}(\hat{\beta}_n - \beta^*))$ to $(\mathbb{V}, z_\infty)$ in $l^\infty(\mathbb{R}^p) \times \mathbb{R}^p$ follows. The next step will be to use the extended continuous mapping theorem (Kosorok 2008). Define the sets

$$\begin{aligned}\mathcal{O} &\triangleq \{\gamma \in \mathbb{R}^p : P1_{X^\intercal\gamma=0} = 0\}\\ \mathcal{S} &\triangleq \{\nu \in l^\infty(\mathbb{R}^p) : \nu \text{ has } ||.|| \text{ continuous sample paths on } \mathcal{O}\}.\end{aligned}$$

Then $(\mathbb{V}, z_\infty) \in (\mathcal{S}, \mathcal{O})$ with probability one. To see this, let $\beta_n \to \beta_0$ in $\mathcal{O}$. If $\aleph \triangleq \{x \in \mathbb{R}^p : x^\intercal\beta_0 = 0\}$ then $P\aleph = 0$. Let $x \in \aleph^c$ and $y \in \{-1, 1\}$, then $1_{yx^\intercal\beta_n<0}1_{x^\intercal\beta^*=0} \to 1_{yx^\intercal\beta_0<0}1_{x^\intercal\beta^*=0}$ and by the dominated convergence theorem

$$P1_{YX^\intercal\beta_n<0}1_{X^\intercal\beta^*=0} \to P1_{YX^\intercal\beta_0<0}1_{X^\intercal\beta^*=0}.$$

The form of the variance-covariance function of $\mathbb{V}$ shows that if $\beta_n \to \beta_0$ in $\mathcal{O}$ then $\rho_\mathbb{V}(\beta_n, \beta_0) \to 0$, where $\rho_\mathbb{V}$ is the covariance semimetric $\rho_\mathbb{V}(s,t) \triangleq P(\mathbb{V}(s) - \mathbb{V}(t))^2$. Therefore, the composition map $q : l^\infty(\mathbb{R}^p) \times \mathbb{R}^p \mapsto \mathbb{R}$ given by $q(u, v) \triangleq u(v)$ is continuous on $\mathcal{S} \times \mathcal{O}$. The extended continuous mapping theorem gives

$$q(\hat{\mathbb{V}}_n, \sqrt{n}(\hat{\beta}_n - \beta^*)) = \sqrt{n}(\mathbb{P}_n - P)1_{X^\intercal\beta^*=0}1_{YX^\intercal\hat{\beta}_n<0} \rightsquigarrow q(\mathbb{V}, z_\infty) = \mathbb{V}(z_\infty).$$

We now consider the second term $\sqrt{n}(\mathbb{P}_n - P)1_{X^\intercal\beta^*\neq0}1_{YX^\intercal\hat{\beta}_n<0}$. The class of functions $\mathcal{G} \triangleq \{1_{X^\intercal\beta^*\neq0}1_{YX^\intercal\gamma<0} : \gamma \in \mathbb{R}^p\}$ is a P-measurable, uniformly bounded, VC class and is hence Donsker. Thus, if we define the process $\hat{\mathbb{B}}_n(\gamma) \triangleq \sqrt{n}(\mathbb{P}_n - P)1_{X^\intercal\beta^*\neq0}1_{YX^\intercal\gamma<0}$ then $\hat{\mathbb{B}}_n \rightsquigarrow \mathbb{B}$ in $l^\infty(\mathbb{R}^p)$. Here, $\mathbb{B}$ is a Brownian-Bridge indexed by $\mathcal{G}$ (or, equivalently $\mathbb{R}^p$). From the form of the variance-covariance function of $\mathbb{B}$, it follows that if $\beta_n \to \beta^*$ then $\rho_\mathbb{B}(\beta_n, \beta^*) \to 0$. Utilizing this result and the equicontinuity of $\hat{\mathbb{B}}_n$ we have $\hat{\mathbb{B}}_n(\hat{\beta}_n) = \hat{\mathbb{B}}_n(\beta^*) + o_{P^*}(1)$. Furthermore, from Slutsky's lemma (see, for example, van der Vaart and Wellner 1996) we have $\hat{\mathbb{B}}_n(\hat{\beta}_n) \rightsquigarrow \mathbb{B}(\beta^*)$.

Finally, joint convergence of $(\hat{\mathbb{V}}_n, \sqrt{n}(\hat{\beta}_n - \beta^*), \hat{\mathbb{B}}_n)$ is established using the Cramer-

Wold device to obtain finite-dimensional convergence (and the fact that the for any $\alpha \in \mathbb{R}^p$ the class $\mathcal{F} + \mathcal{G} + \alpha^\intercal g(X, Y, \beta^*)$ is Donsker). The result follows from the extended continuous mapping theorem applied twice. $\square$

Proof. [Corollary 3.2 part (2)] The class of functions

$$\mathcal{H} \triangleq \left\{ 1_{YX^\intercal u<0} 1_{\frac{(X^\intercal \gamma)^2}{X^\intercal \Sigma X} \le \alpha} : u, \gamma \in \mathbb{R}^p, \alpha \in \mathbb{R}_+ \right\}$$

is a P-measurable, uniformly bounded VC class of functions and hence is Donsker. Thus, if we define the process $\hat{\mathbb{W}}_n(u, \gamma, \alpha) \triangleq \sqrt{n}(\mathbb{P}_n - P) 1_{YX^\intercal u<0} 1_{\frac{(X^\intercal \gamma)^2}{X^\intercal \Sigma X} \le \alpha}$ then $\hat{\mathbb{W}}_n \rightsquigarrow \mathbb{W}$ in $l^\infty(\mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}_+)$. Where $\mathbb{W}$ is a Brownian-Bridge indexed by $\mathcal{H}$ (or equivalently, $\mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}_+$). Notice that $\mathbb{W}(u, \beta^*, 0) =_{\mathcal{L}} \mathbb{V}(u)$ as a process indexed by $u \in \mathbb{R}^p$. Let $\beta_n \to \beta^*$ in $\mathbb{R}^p$, $\epsilon_n \to 0$ in $\mathbb{R}_+$ satisfy $\epsilon_n^{-1} ||\beta_n - \beta^*||^2 \to 0$, then examination of the form of the variance-covariance function of $\mathbb{W}$ shows that $\sup_{u \in \mathbb{R}^p} \rho_{\mathbb{W}} \left((u, \beta_n, \epsilon_n), (u, \beta^*, 0) \right) \to 0$. This smoothness of $\rho_{\mathbb{W}}$ combined with the equicontinuity of $\hat{\mathbb{W}}_n$ implies that $\hat{\mathbb{W}}_n(u, \hat{\beta}_n, 1/a_n) = \hat{\mathbb{W}}_n(u, \beta^*, 0) + o_{P^*}(1)$ where the remainder term goes to zero in outer probability *uniformly* over u. Thus by Slutsky's theorem $\hat{\mathbb{W}}_n(u, \hat{\beta}_n, 1/a_n) \rightsquigarrow \mathbb{W}(u, \beta^*, 0) =_{\mathcal{L}} \mathbb{V}(u)$. We have shown

$$\sqrt{n}(\mathbb{P}_n - P) 1_{YX^\intercal u<0} 1_{\frac{(X^\intercal \hat{\beta}_n)^2}{X^\intercal \Sigma X} \le \frac{1}{a_n}} \rightsquigarrow \mathbb{V}(u)$$

in $l^\infty(\mathbb{R}^p)$.

We now turn attention to the limiting distribution of $\sqrt{n}(\mathbb{P}_n - P) 1_{YX^\intercal \hat{\beta}_n<0} 1_{\frac{(X^\intercal \hat{\beta}_n)^2}{X^\intercal \Sigma X} > \frac{1}{a_n}}$. The class of functions $\mathcal{Q} \triangleq \left\{ 1_{YX^\intercal \gamma<0} 1_{\frac{(X^\intercal \gamma)^2}{X^\intercal \Sigma X} > \alpha} : \gamma \in \mathbb{R}^p, \alpha \in \mathbb{R}_+ \right\}$ is a P-measurable, uniformly bounded, VC class of functions and hence is Donsker. Thus, if we define $\hat{\mathbb{D}}_n(\gamma, \alpha) \triangleq \sqrt{n}(\mathbb{P}_n - P) 1_{YX^\intercal \gamma<0} 1_{\frac{(X^\intercal \gamma)^2}{X^\intercal \Sigma X} > \alpha}$ then $\hat{\mathbb{D}}_n \rightsquigarrow \mathbb{D}$ in $l^\infty(\mathbb{R}^p \times \mathbb{R}_+)$. In parallel with the first portion of this proof, suppose that $\beta_n \to \beta^*$ in $\mathbb{R}^p$, and $\epsilon_n \to 0$

in $\mathbb{R}_+$ so that $\epsilon_n^{-1}||\beta_n - \beta^*||^2 \to 0$. Then for any $(x, y) \in \mathbb{R}^p \times \{-1, 1\}$ we have $1_{yx^\intercal\beta_n<0}1_{\frac{(x^\intercal\beta_n)^2}{x^\intercal\Sigma x}>\epsilon_n} \to 1_{yx^\intercal\beta^*<0}1_{x^\intercal\beta^*\neq 0}$ so that from the dominated convergence theorem we have $P1_{YX^\intercal\beta_n<0}1_{\frac{(X^\intercal\beta_n)^2}{X^\intercal\Sigma X}>\epsilon_n} \to P1_{YX^\intercal\beta^*<0}1_{X^\intercal\beta^*\neq 0}$. The form of the variance-covariance function of $\mathbb{D}$ shows that $\rho_{\mathbb{D}}\left((\beta_n, \epsilon_n), (\beta^*, 0)\right) \to 0$ when β_n and ϵ_n satisfy the conditions given above. Using an argument identical to that in the first portion of this proof it follows that $\hat{\mathbb{D}}_n(\hat{\beta}_n, 1/a_n) = \hat{\mathbb{D}}_n(\beta^*, 0) + o_{P^*}(1)$. Applying Slutsky's theorem shows $\hat{\mathbb{D}}_n(\hat{\beta}_n, 1/a_n) \rightsquigarrow \mathbb{D}(\beta^*, 0) =_{\mathcal{L}} \mathbb{B}(\beta^*)$.

Finally, joint convergence of $(\hat{\mathbb{W}}_n, \hat{\mathbb{D}}_n)$ in $l^\infty(\mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}_+) \times l^\infty(\mathbb{R}^p \times \mathbb{R}_+)$ is established using the Cramer-Wold device to obtain finite dimensional convergence (utilizing the fact that $\mathcal{H} + \mathcal{Q}$ is Donsker). Asymptotic tightness and measurabiity in the product space follows from tightness and measurability of the marginals (Kosorok 2008). The continuous mapping theorem gives the result. □

Proof. [Corollary 3.4] This proof parallels its population analogue given above. We assume without loss of generality that $\hat{\Sigma}_n$ is positive definite for all n with probability one. As a first step we derive the limiting distribution of $\sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)1_{YX^\intercal u<0}1_{\frac{(X^\intercal\hat{\beta}_n^{(b)})^2}{X^\intercal\hat{\Sigma}_n X}\leq\frac{1}{a_n}}$. The class of functions $\mathcal{I} \triangleq \left\{1_{YX^\intercal u<0}1_{\frac{(X^\intercal\gamma)^2}{X^\intercal\Omega X}\leq\alpha} : u, \gamma \in \mathbb{R}^p, \Omega \in PD(\mathbb{R}^{p\times p}), \alpha \in \mathbb{R}_+\right\}$ where $PD(\mathbb{R}^{p\times p})$ is the space of real-valued $p \times p$ positive definite matrices, is a P-measurable, uniformly bounded, VC class of functions and hence Donsker. If we define $\hat{\mathbb{Z}}_n^{(b)}(u, \gamma, \Omega, \alpha) \triangleq \sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)1_{YX^\intercal u<0}1_{\frac{(X^\intercal\gamma)^2}{X^\intercal\Omega X}\leq\alpha}$ then $\hat{\mathbb{Z}}_n^{(b)} \rightsquigarrow \mathbb{Z}$ almost surely in $l^\infty(\mathbb{R}^p \times \mathbb{R}^p \times PD(\mathbb{R}^{p\times p}) \times \mathbb{R}_+)$ (van der Vaart and Wellner 1996). Where $\mathbb{Z}$ is a Brownian Bridge. Let $\beta_n \to \beta^*$ in $\mathbb{R}^p$, and $\epsilon_n \to 0$ in $\mathbb{R}_+$ satisfy $\epsilon_n^{-1}||\beta_n - \beta^*||^2 \to 0$, in addition, suppose that $\Sigma_n \to \Sigma$ in $PD(\mathbb{R}^{p\times p})$. Then, an argument similar to the one given in the proof of Corollary 3.2 shows that $\hat{\mathbb{Z}}_n^{(b)}(u, \hat{\beta}_n^{(b)}, \hat{\Sigma}_n, 1/a_n) = \hat{\mathbb{Z}}_n^{(b)}(u, \beta^*, \Sigma, 0) + o_{P_M}(1)$ in probability, note that the error term does not depend on u. Thus by Slutsky's theorem $\hat{\mathbb{Z}}_n^{(b)}(u, \hat{\beta}_n^{(b)}, \hat{\Sigma}_n, 1/a_n)$ converges weakly in probability to $\mathbb{Z}(u, \beta^*, \Sigma, 0) =_{\mathcal{L}} \mathbb{V}(u)$.

We now establish the limiting distribution of the second term

$$\sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)1_{YX^\intercal\hat{\beta}_n^{(b)}<0}1_{\frac{(X^\intercal\hat{\beta}_n^{(b)})^2}{X^\intercal\hat{\Sigma}_nX}>\frac{1}{a_n}}.$$

The class of functions $\mathcal{E} \triangleq \left\{1_{YX^\intercal\gamma<0}1_{\frac{(X^\intercal\gamma)^2}{X^\intercal\Omega X}>\alpha} : \gamma \in \mathbb{R}^p, \Omega \in PD(\mathbb{R}^{p\times p}), \alpha \in \mathbb{R}_+\right\}$ is a P-measurable, uniformly bounded, VC class of functions and hence Donsker. Thus, if we define, $\hat{\mathbb{J}}_n^{(b)} \triangleq \sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)1_{YX^\intercal\gamma<0}1_{\frac{(X^\intercal\gamma)^2}{X^\intercal\Omega X}>\alpha}$ then $\hat{\mathbb{J}}_n^{(b)} \rightsquigarrow \mathbb{J}$ almost surely in $l^\infty(\mathbb{R}^p \times PD(\mathbb{R}^{p\times p}) \times \mathbb{R}_+)$. Where $\mathbb{J}$ is a Brownian-Bridge. Again, let $\beta_n \to \beta^*$ in $\mathbb{R}^p$, and $\epsilon_n \to 0$ in $\mathbb{R}_+$ so that $\epsilon_n^{-1}||\beta_n - \beta^*||^2 \to 0$, further let $\Sigma_n \to \Sigma$ in $PD(\mathbb{R}^{p\times p})$ then for any $(x, y) \in \mathbb{R}^p \times \{-1, 1\}$ we have $1_{yx^\intercal\beta_n}1_{\frac{(x^\intercal\beta_n)^2}{X^\intercal\Sigma X}>\epsilon_n} \to 1_{yx^\intercal\beta^*<0}1_{x^\intercal\beta^*\neq 0}$. The dominated convergence theorem gives $P1_{yx^\intercal\beta_n}1_{\frac{(x^\intercal\beta_n)^2}{X^\intercal\Sigma X}>\epsilon_n} \to P1_{yx^\intercal\beta^*<0}1_{x^\intercal\beta^*\neq 0}$. The form of the variance-covariance function of $\mathbb{J}$ we have that $\rho_{\mathbb{J}}((\beta_n, \Sigma_n, \epsilon_n), (\beta^*, \Sigma, 0)) \to 0$ for β_n, ϵ_n, and Σ_n satisfying the above assumptions. This smoothness in $\rho_{\mathbb{J}}$ and the equicontinuity of $\hat{\mathbb{B}}_n^{(b)}$ shows that $\hat{\mathbb{B}}_n^{(b)}(\hat{\beta}_n^{(b)}, \hat{\Sigma}_n, 1/a_n) = \hat{\mathbb{B}}_n^{(b)}(\beta^*, \Sigma, 0) + o_{P_M}(1)$. Thus, $\hat{\mathbb{J}}_n^{(b)}(\hat{\beta}_n^{(b)}, \hat{\Sigma}_n, 1/a_n)$ converges weakly to $\mathbb{J}(\beta^*, \Sigma, 0) =_{\mathcal{L}} \mathbb{B}(\beta^*)$ in probability.

Finally, joint convergence of the finite dimensional distributions of $(\hat{\mathbb{Z}}_n^{(b)}, \hat{\mathbb{J}}_n^{(b)})$ is established by means of the Cramer-Wold device. Asymptotic tightness and measurability follow from tightness and measurability of the marginals (Kosorok 2008). The continuous mapping theorem gives the result noting that $d^*_{BL_1(\mathbb{R})}$ metrizes convergence. □

Proof. [Theorem 3.5] The proof of part (1) follows from the same steps as the proof of Corollary 3.2. The only difference is to notice that $\sqrt{n}(\hat{\beta}_n - \beta_n^*) = -H^{-1}\sqrt{n}(\mathbb{P}_n - P_n)g(X, Y, \beta^*) + o_P(1)$ so that $\sqrt{n}(\hat{\beta}_n - \beta_n^*) \rightsquigarrow z_\infty =_{\mathcal{L}} N(0, H^{-1}\Omega H^{-1})$ and in turn we have $\sqrt{n}(\hat{\beta}_n - \beta^*) = \sqrt{n}(\hat{\beta}_n - \beta_n^*) + \gamma \rightsquigarrow z_\infty + \gamma =_{\mathcal{L}} N(\gamma, H^{-1}\Omega H^{-1})$. The proof of part (2) follows exactly the same steps as in the proof of Theorem 3.1 and is so omitted. □

CHAPTER III

Statistical inference for dynamic treatment regimes

Dynamic treatment regimes, also known as *treatment policies*, are increasingly being used to operationalize sequential clinical decision making associated with patient care. Common approaches to constructing a dynamic treatment regime from data, such as Q-learning, employ non-smooth functionals of the data. Therefore, simple inferential tasks such as constructing a confidence interval for the parameters in the Q-function are complicated by non-regular asymptotics under certain commonly-encountered generative models. Methods that ignore this non-regularity can suffer from poor performance in small samples. We construct confidence intervals for the parameters in the Q-function by constructing smooth, data-dependent, upper and lower bounds on these parameters and then applying the bootstrap. The confidence interval is adaptive in that although it is conservative for non-regular generative models, it achieves asymptotically exact coverage elsewhere. The small sample performance of the method is evaluated on a series of examples and compares favorably to previously published competitors. Finally, we illustrate the method using data from the Adaptive Interventions for Children with ADHD study (Pelham and Fabiano 2008).

This work is motivated by our involvement in the Adaptive Interventions for Children with Attention Deficit Hyperactivity Disorder (ADHD) study (Pelham and Fabiano 2008). ADHD affects an estimated 5%-10% of school aged children, and is characterized by inattention, hyperactivity, and impulsivity (Pliszka 2007). In the years preceding the study, clinicians debated the comparative effectiveness of behavioral modification therapy versus medication as treatment options for ADHD (Pliszka 2007; Pelham and Fabiano 2008). As a consequence, a SMART trial was conducted with the general aim of estimating the treatment policy that achieves the greatest reduction of ADHD symptoms among school age children. This SMART study is composed of two stages. In the first stage, children were randomized with equal probability into one of two treatment groups (low-dose behavioral modification therapy, low-dose medication). After a burn-in period of eight weeks, children were evaluated monthly and at each evaluation deemed either a responder or non-responder. (The operationalized definition of nonresponse is given in the trial protocol which can be found in (Pelham and Fabiano 2008).) Non-responders were immediately re-randomized to either (i) augmentation of treatment, so that the child received both low-dose medication *and* low-dose behavioral modification therapy, or (ii) intensification of treatment, so that the child received an increased dosage of their current (stage one) treatment. Responders were not re-randomized and continued to receive their current treatment at the current dosage level.

Data collected in a SMART trial, like the ADHD study, can be used to estimate an optimal treatment policy. This estimation typically uses an extension of regression to multistage decision making problems. The extension we consider in this paper is the Q-learning algorithm (Watkins 1989, Murphy 2005). A variety of other extensions exist in the statistical literature (Murphy 2003, Robins 2004, Blatt et al. 2004, Moodie et al. 2007, Henderson et al. 2009, Zhao et al. 2009). However *all of these extensions* suffer from the same problem of non-regularity that we focus on in this paper (Robins 2004; Moodie and Richardson 2007; Henderson et al. 2009; Chakraborty et al. 2009; Moodie et al. 2010).

In this paper we provide a method for constructing confidence intervals for parameters arising in the Q-Learning algorithm. The primary challenge to this task is that the estimators are non-smooth functionals of the data—in particular, the formula for the estimators involves the use of the max operator, which is non-differentiable. Robins (2004) notes two problems resulting from the non-differentiability of the max operator. First while the estimators of the regression coefficients are consistent, their limiting distributions can have nonzero mean; that is, there is estimation bias on the order of $1/\sqrt{n}$ for some generative models. Second the regression coefficient estimators are non-regular (Bickel 2001; Tsiatis 2006). That is, the limiting distribution changes abruptly as one smoothly varies the underlying generative model and as a practical consequence, implies that common approaches based on the bootstrap and Taylor series arguments provide inconsistent interval estimators and can behave poorly in small samples (Andrews and Ploberger 1994; Andrews 2001, 2002; Leeb and Poetscher 2005) To deal with the non-regularity, Robins (2004) proposes conducting inference via the use of projection confidence intervals. This method, while yielding consistent confidence sets, is computationally difficult due to the need to solve a very difficult non-convex optimization problem and may result in a confidence set that is the union of intervals. More importantly, the resultant set, when estimated numerically, tends to be much too large to be useful in small samples.

The adaptive confidence interval proposed here is based on smooth, data-dependent, upper and lower bounds on the estimators involved in the regression models used by Q-learning. Confidence intervals are formed by bootstrapping these bounds. The proposed confidence interval is adaptive in that although it is conservative for non-regular generative models, it achieves asymptotically exact coverage elsewhere. The method requires

Other authors have focused on reducing the bias of order $1/\sqrt{n}$ discussed above (recall the estimators of the regression coefficients can have limiting distributions with nonzero means, thus bias). The methods of Moodie and Richardson (2007), Chakraborty et al. (2009) and Song et al. (2010) reduce the estimation bias via the use of thresholding. As is well-known, the use of thresholding (or penalization that induces induce variable selection) leads to non-regular estimation (see Lee 2003; Leeb and Potscher 2005 and references therein; Chatterjee and Lahiri 2009, 2010). The soft-thresholding method proposed by Chakraborty et al. applies to problems with two stages and binary treatments, and effectively reduces the bias as demonstrated by good simulation performance. Unfortunately, the method is provably inconsistent. Song et al. use a lasso-like penalization; here appropriate choice of the tuning parameter leads to consistent estimation. Neither method has been generalized to handle more than two treatments. The method proposed by Moodie and Richardson is quite similar to that proposed by Chakraborty et al.; the primary difference being the use of a hard-threshold. In Chakraborty et al. confidence intervals are constructed by use of the bootstrap whereas in Song et al., confidence intervals are produced via Taylor series arguments. Both methods work well in the simulations provided. However, the standard bootstrap is inconsistent in non-regular settings (Shao 1994; Beran 1997) and confidence intervals based on Taylor series arguments do not capture the variation due to variable selection/thresholding (cite Leeb).

We do not alter the Q-Learning algorithm and thus do not attempt to reduce the bias of order $1/\sqrt{n}$ in the estimators of the regression coefficients. We do this for several reasons: first it is known that in settings in which there is no unbiased estimator, attempts to eliminate the bias for some parameter values must lead to large mean square error for other parameter values (Doss and Sethuraman, 1989; Brown and Liu 1993; Chen 2004; Hirano and Portor 2009) and second, toy simulations provided in the supplementary material support this finding. See the discussion for further comments.

An overview of the remainder of this paper is as follows: Section 2 introduces the problem of constructing confidence intervals under the simplest possible setting, in which there are two stages of treatment and two treatments available at each stage. The adaptive confidence interval (ACI) is introduced and asymptotic properties are provided. Section 3 generalizes the problem, our approach, and the theoretical results to the class of problems with two stages of treatment and an arbitrary number of treatments at each stage. In Section 4, we provide an empirical comparison of the ACI with the bootstrap and the use of thresholding as represented in Chakraborty et al. (2009) on a number of test cases. The ACI compares favorably with these competitors. Section 5 contains an application of the ACI to the analysis of the ADHD study and a discussion of future work. An extension of the ACI to an arbitrary number of stages of treatment, and an arbitrary number of treatments at each stage, is given in the supplementary material.

3.1 Two stages of binary treatment

In this section, we introduce the problem of constructing a confidence interval for the parameters in the Q-function when there are two stages of treatment and two treatments are available at each stage. First, we provide the requisite notation and review the Q-learning algorithm. Second, we introduce the adaptive confidence interval (ACI) for the parameters defining the regression models used in Q-learning. We prove that the ACI provides asymptotically valid confidence intervals under both fixed and local alternatives. Since the ACI involves the use of a tuning parameter, we empirically examine the robustness of the ACI to the choice of this tuning parameter.

We use uppercase letters such as X and A to denote random variables, and lowercase letters such as x and a to denote instances of these random variables. The data consist of n trajectories drawn *i.i.d.* from some fixed and unknown distribution P. Each trajectory $(X_1, A_1, Y_1, X_2, A_2, Y_2)$ is a sequence of random variables collected at two stages $t = 1, 2$; $X_t \in \mathbb{R}^{p_t}$ denotes patient measurements collected prior to the tth assignment of treatment, $A_t \in \{-1, 1\}$, denotes the binary treatment (also called an action) assigned at stage t and $Y_t \in \mathbb{R}$ is a measure of patient response following the assignment of treatment at stage t. We assume that Y_t has been coded so that a higher value corresponds to a better clinical outcome. Let $H_t = \{X_1, A_1, \ldots, X_t\}$ be the patient history, e.g., the information available to the decision maker *before* the assignment of the tth treatment A_t. Furthermore, we assume that the treatments, A_t, are randomly assigned to patients at each stage with probabilities possibly depending on patient history.

We wish to use data like the above to inform the construction of a Dynamic Treatment Regime (DTR). A DTR is sequence of decision rules, one for each stage of treatment, that takes as input the patient history and gives as output a recommended treatment. More formally, a DTR $\pi = (\pi_1, \pi_2)$ is an ordered pair of functions π_t so that $\pi_t : \mathcal{H}_t \mapsto \{-1, 1\}$ where $\mathcal{H}_t \subseteq \mathbb{R}^{d_t}$ is the domain of H_t. Let $\mathbb{E}^\pi$ denote the joint expectation over H_t, A_t, Y_t for $t = 1, 2$ under the restriction that $A_t = \pi(H_t)$. The objective is to learn a DTR π which comes close to maximizing the expected clinical outcome $\mathbb{E}^\pi(Y_1 + Y_2)$. One way to estimate an optimal DTR is using the Q-learning algorithm (Watkins 1989), which can be conceptualized as an extension of regression to multistage decision making. More precisely, Q-learning is a form of approximate dynamic programming, where the conditional mean responses are estimated from the data since they cannot be computed explicitly. We now describe the Q-learning algorithm with function approximation as in (Murphy 2005). To start, define

$$\begin{aligned} Q_2(h_2, a_2) &\triangleq \mathbb{E}\left(Y_2 \middle| H_2 = h_2, A_2 = a_2\right) \qquad (3.1)\\ Q_1(h_1, a_1) &\triangleq \mathbb{E}\left(Y_1 + \max_{a_2 \in \{-1,1\}} Q_2(H_2, a_2) \middle| H_1 = h_1, A_1 = a_1\right); \qquad (3.2)\end{aligned}$$

the functions $Q_t(h_t, a_t)$ $t = 1, 2$ are known as Q-functions. At each stage of treatment t the Q-function reflects the quality (hence the letter “Q”) of the treatment a_t given the patient history h_t. If the conditional expectations in the preceding display were known, then dynamic programming provides an optimal DTR given by $\pi^{dp}(h_t) \triangleq \arg\max_{a_t \in \{-1,1\}} Q_t(h_t, a_t)$. In most practical settings this mean function must be approximated from data. In this paper we consider linear approximations to the conditional mean function. Specifically, we employ a model of the form

$$Q_t(h_t, a_t; \beta_t) = \beta_{t,0}^\intercal h_{t,0} + \beta_{t,1}^\intercal h_{t,1} 1_{a_t=1}, \qquad (3.3)$$

where $h_{t,0}$ and $h_{t,1}$ are vectors of features comprising the patient history. Note that according to the model, if $h_{t,1}^\intercal \beta_{t,1} \approx 0$ then both treatments $a_t = 1$ and $a_t = -1$ yield the approximately same response for a patient with history $H_{t,1} = h_{t,1}$. That is, that there is *not* a unique best treatment for a patient with history $H_{t,1} = h_{t,1}$. Conversely, if $|h_{2,1}^\intercal \beta_{t,1}| \gg 0$ then exactly one treatment yields the best expected outcome for a patient with history $H_{t,1} = h_{t,1}$. We use β_t to denote $(\beta_{t,0}^\intercal, \beta_{t,1}^\intercal)^\intercal$. Let $\mathbb{P}_n$ denote the empirical measure. The Q-learning algorithm proceeds as follows:

1. Regress Y_2 on H_2 and A_2 using (3.3) to obtain

$$\hat{\beta}_2 \triangleq \arg\min_{\beta_2} \mathbb{P}_n \left(Y_2 - Q_2(H_2, A_2; \beta_2)\right)^2,$$

and subsequently the approximation $Q_2(h_2, a_2; \hat{\beta}_2)$ to the conditional mean $Q_2(h_2, a_2)$.

2. (a) Define the predicted future reward following the optimal policy as:

$$\begin{aligned} \tilde{Y}_1 &\triangleq Y_1 + \max_{a_2 \in \{-1,1\}} Q_2(H_2, A_2; \hat{\beta}_2) && (3.4) \\ &= Y_1 + H_{2,0}^{\intercal}\hat{\beta}_{2,0} + \left[H_{2,1}^{\intercal}\hat{\beta}_{2,1} \right]_{+}, && (3.5) \end{aligned}$$

   where $[z]_+$ denotes the positive part of z.

   (b) Regress $\tilde{Y}_t$ on H_1 and A_1 using (3.3) to obtain $\hat{\beta}_1 \triangleq \arg\min_{\beta_1} \mathbb{P}_n(\tilde{Y}_1 - Q_1(H_1, A_1; \beta_1))^2$.

3. Define the estimated optimal DTR as $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2)$ so that

$$\hat{\pi}_t(h_t) \triangleq \arg \max_{a_t \in \{-1,1\}} Q_t(h_t, a_t; \hat{\beta}_t).$$

Examination of the above procedure make apparent the close connection between Q-learning and dynamic programming. For further elaboration see (Watkins 1989; Murphy 2005).

The second stage population coefficients, β_2^* satisfy $\beta_2^* \triangleq \arg\min_{\beta_2} P\left(Y_2 - Q_2(H_2,\ A_2; \beta_2)\right)$. Define $\tilde{Y}_1^* \triangleq Y_1 + H_{2,0}^{\intercal}\beta_{2,0}^* + \left[H_{2,1}^{\intercal}\beta_{2,1}^*\right]_+$, then the first stage population coefficients β_1^* are given by

$$\beta_1^* \triangleq \arg\min_{\beta_1} P\left(\tilde{Y}_1^* - Q_1(H_1, A_1; \beta_1)\right)^2.$$

The goal of this paper is the development of asymptotically valid confidence intervals for first stage coefficients, β_1^*. Note the construction of confidence intervals for the second stage coefficients is standard.

The ACI is formed by constructing smooth data-dependent bounds on $c^{\intercal}\sqrt{n}(\hat{\beta}_1 -$

β_1^*). As a first step toward developing the bounds used in the ACI, we provide a useful decomposition of $c^\intercal\sqrt{n}(\hat{\beta}_1-\beta_1^*)$. Define $B_1 \triangleq (H_{1,0}^\intercal, H_{1,1}^\intercal 1_{A_1=1})$ so that instances of B_1 form the rows of the design matrix in the first stage regression. Let $\Sigma_{1,n} \triangleq \mathbb{P}_n B_1^\intercal B_1$, then examination of the normal equations shows that $\hat{\beta}_1 = \Sigma_{1,n}^{-1}\mathbb{P}_n B_1^\intercal \tilde{Y}_1$. Hence, for any $c \in \mathbb{R}^{\dim(\beta_1^*)}$ it follows that $c^\intercal\sqrt{n}(\hat{\beta}_1-\beta_1^*) = c^\intercal\Sigma_{1,n}^{-1}\sqrt{n}\mathbb{P}_n B_1^\intercal\left(\tilde{Y}_1 - B_1\beta_1^*\right)$, which, using the definition of $\tilde{Y}_1$, can be further decomposed as

$$c^\intercal\mathbb{W}_n + c^\intercal\Sigma_{1,n}^{-1}\mathbb{P}_n B_1^\intercal\mathbb{U}_n, \tag{3.6}$$

where

$$\begin{aligned}
\mathbb{W}_n &= \Sigma_{1,n}^{-1}\sqrt{n}\mathbb{P}_n B_1^\intercal\Big[\left(Y_1 + H_{2,0}^\intercal\beta_{2,0}^* + \left[H_{2,1}^\intercal\beta_{2,1}^*\right]_+ - B_1\beta_1^*\right) + H_{2,0}^\intercal\left(\hat{\beta}_{2,0}-\beta_{2,0}^*\right)\Big],\\
\mathbb{U}_n &= \sqrt{n}\left(\left[H_{2,1}^\intercal\hat{\beta}_{2,1}\right]_+ - \left[H_{2,1}^\intercal\beta_{2,1}^*\right]_+\right).
\end{aligned}$$

The second term in (3.6) is non-smooth which can be seen from the definition of $\mathbb{U}_n$. To illustrate the effect of this non-smoothness, fix $H_{2,1} = h_{2,1}$. If $h_{2,1}^\intercal\beta_{2,1}^* > 0$, then $\mathbb{U}_n\big|_{H_{2,1}=h_{2,1}}$ is easily seen to be asymptotically normal with mean zero. On the other hand, if $h_{2,1}^\intercal\beta_{2,1}^* = 0$, then $\mathbb{U}_n\big|_{H_{2,1}=h_{2,1}} = \left[h_{2,1}^\intercal\sqrt{n}(\hat{\beta}_{2,1}-\beta_{2,1})\right]_+$ which converges weakly to positive part of a mean zero normal random variable. Thus, the limiting distribution $c^\intercal\sqrt{n}(\hat{\beta}_1-\beta_1^*)$ depends abruptly on both the true parameter $\beta_{2,1}^*$ and the distribution of patient features $H_{2,1}$. In particular, the limiting distribution of $c^\intercal\sqrt{n}(\hat{\beta}_1-\beta_1^*)$ depends on the frequency of patient features $H_{2,1} = h_{2,1}$ for which there is no treatment effect (e.g. $h_{2,1}^\intercal\beta_{2,1}^* = 0$). We construct upper and lower bound on $c^\intercal\sqrt{n}(\hat{\beta}_1-\beta_1^*)$ by first partitioning the data into two sets (i) patients for which there appears to be a treatment effect, and (ii) patients where it appears there is no treatment effect. The bounds are formed by bounding the error of the overall approximation due to misclassification of patients during the partitioning.

The idea of conducting a preliminary hypothesis test prior to forming estimators or confidence intervals is known as preliminary testing or pretesting (see Olshen 1973); indeed estimators formed by thresholding implicitly use a pretest. Pretesting has been used in Econometrics to provide hypothesis tests and confidence intervals in nonregular settings (Andrews 2001, Andrews & Soares 2007; Andrews & Guggenberger 2009, Cheng 2008). In these settings one can identify a small (usually one) number of parameters values at which non-regularity occurs. If the pretest rejects, a standard critical value is used to form confidence interval; if the pretest accepts, the maximal critical value over all possible local alternatives is used to form the confidence interval. In this paper the situation is somewhat different as non-regularity occurs for any combination of the distribution of the $H_{2,1}$ and $\beta_{2,1}$ for which $P[H_{2,1}^T\beta_{2,1} = 0] > 0$. Thus we take a different tactic from that employed in the Econometrics literature. Instead, we conduct a pretest for each individual in the data set as follows. Define $\mathbb{V}_n \triangleq \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1}^*)$, and $\Sigma_{2,n}^{(2,2)}$ to be the plug-in estimator of the asymptotic covariance matrix of $\mathbb{V}_n$. In addition, let $\Xi_{2,n}$ denote the inverse of the matrix square-root of $\Sigma_{2,n}^{(2,2)}$. The pretest is based on $T_n(h_{2,1}) \triangleq \frac{n(h_{2,1}^\intercal\hat{\beta}_{2,1})^2}{h_{2,1}^\intercal\Sigma_{2,n}^{(2,2)}h_{2,1}}$; note that $T_n(h_{2,1})$ corresponds to the usual test statistic when testing the null hypothesis, $h_{2,1}^\intercal\beta_{2,1}^* = 0$.

The upper bound on $c^\intercal\sqrt{n}(\hat{\beta}_1 - \beta_1^*)$ is given by

$$\begin{aligned}\mathcal{U}(c) &\triangleq c^\intercal\mathbb{W}_n + c^\intercal\Sigma_{1,n}^{-1}\mathbb{P}_nB_1^\intercal\mathbb{U}_n1_{T_n(H_{2,1})>\lambda_n} + c^\intercal\Sigma_{1,n}^{-1}\mathbb{P}_nB_1^\intercal\left[H_{2,1}^\intercal\mathbb{V}_n\right]_+1_{T_n(H_{2,1})\le\lambda_n}\\ &+ \sup_{\gamma\in\mathcal{S}_n} c^\intercal\Sigma_{1,n}^{-1}\mathbb{P}_nB_1^\intercal\Big(\left[H_{2,1}^\intercal(\mathbb{V}_n+\gamma)\right]_+ - \left[H_{2,1}^\intercal\gamma\right]_+ - \left[H_{2,1}^\intercal\mathbb{V}_n\right]_+\Big)1_{T_n(H_{2,1})\le\lambda_n}, \quad (3.7)\end{aligned}$$

where $\mathcal{S}_n \triangleq \left\{\gamma \in \mathbb{R}^{\dim(\beta_{2,1}^*)} : ||\Xi_{2,n}(\gamma - \sqrt{n}\beta_{2,1}^*)||_\infty^2 \le \lambda_n\right\}$, and λ_n is a tuning parameter that we discuss in detail below. A lower bound $\mathcal{L}(c)$ can be defined by replacing the sup with an inf. The intuition behind this upper bound is as follows. Notice that the second term, $c^{\intercal}\Sigma_{1,n}^{-1}\mathbb{P}_n B_1^{\intercal}\mathbb{U}_n$, in (3.6) is equal to

$$c^{\intercal}\Sigma_{1,n}^{-1}\mathbb{P}_n B_1^{\intercal}\mathbb{U}_n 1_{T_n(H_{2,1})>\lambda_n} + c^{\intercal}\Sigma_{1,n}^{-1}\mathbb{P}_n B_1^{\intercal}\left[H_{2,1}^{\intercal}\mathbb{V}_n\right]_+ 1_{T_n(H_{2,1})\leq\lambda_n} \\ + c^{\intercal}\Sigma_{1,n}^{-1}\mathbb{P}_n B_1^{\intercal}\left(\mathbb{U}_n - \left[H_{2,1}^{\intercal}\mathbb{V}_n\right]_+\right) 1_{T_n(H_{2,1})\leq\lambda_n}. \quad (3.8)$$

The first and second terms in (3.8) correspond to differing approximations according to the partitioning of the data. The first term in (3.8) involving $\mathbb{U}_n$ is left unchanged on the partition where $T_n(h_{2,1})$ is large and thus there appears to be a large treatment effect. Since $\mathbb{U}_n\big|_{H_{2,1}=h_{2,1}} = \left[h_{2,1}^{\intercal}\mathbb{V}_n\right]_+$ when $h_{2,1}^{\intercal}\beta_{2,1}^{*} = 0$, we replace the term $\mathbb{U}_n$ by $[h_{2,1}^{\intercal}\mathbb{V}_n]_+$ when $T_n(h_{2,1})$ is small to obtain the second term in (3.8). The third term above corresponds to the error in the partition (i.e., this term would have been absent had we been able to partition the data according to the unknown $1_{H_{2,1}^{\intercal}\beta_{2,1}^{*}=0}$ instead of $1_{T_n(H_{2,1})\leq\lambda_n}$). Using the definition of $\mathbb{U}_n$ and $\mathbb{V}_n$, it can be shown that the third term in (3.8) is equal to

$$c^{\intercal}\Sigma_{1,n}^{-1}\mathbb{P}_n B_1^{\intercal}\left(\left[H_{2,1}^{\intercal}(\mathbb{V}_n + \sqrt{n}\beta_{2,1}^{*})\right]_+ - \left[H_{2,1}^{\intercal}\sqrt{n}\beta_{2,1}^{*}\right]_+ - \left[H_{2,1}^{\intercal}\mathbb{V}_n\right]_+\right) 1_{T_n(H_{2,1})\leq\lambda_n}. \quad (3.9)$$

Replacing $\sqrt{n}\beta_{2,1}^{*}$ with γ and taking the supremum over any subset of $\mathbb{R}^{\dim(\beta_{2,1}^{*})}$ which contains $\sqrt{n}\beta_{2,1}^{*}$ yields an upper bound on the last term in (3.8). Justififcation of our particular choice $\mathcal{S}_n$ is given below. Combining this result with (3.6) and (3.8) yields (3.7). Theorem 2.1 below provides the asymptotic distribution of (3.7).

The quantity $\mathcal{U}(c)$ forms an upper bound of $c^{\intercal}\sqrt{n}(\hat{\beta}_1 - \beta_1^{*})$ whenever $\sqrt{n}\beta_{2,1}^{*}$ belongs to the set $\mathcal{S}_n$. A potentially important concern is the choice of $\mathcal{S}_n$. Let $\mathcal{H}$ denote the hyperplane in $\mathbb{R}^{\dim(\beta_{2,1}^{*})}$ defined by the normal vector $\beta_{2,1}^{*}$. Let $\mathbb{P}_{\mathcal{H}}$ denote the orthogonal projection onto $\mathcal{H}$. It is proved in the supplement that $\sup_{h_{2,1}\in\mathcal{H}} T_n(h_{2,1})/\lambda_n =$

$o_P(1)$. Using the form of $T_n(h_{2,1})$ we see that this result is equivalent to

$$\sup_{h_{2,1}\in\mathcal{H}} \frac{(h_{2,1}^{\intercal}(\mathbb{V}_n + \sqrt{n}\beta_{2,1}^*))^2}{\lambda_n ||h_{2,1}||^2} = o_P(1).$$

It is natural then to require that any sequence of vectors γ_n in $\mathbb{R}^p$ satisfying

$$\sup_{h_{2,1}\in\mathcal{H}} \frac{(h_{2,1}^{\intercal}(\mathbb{V}_n + \gamma_n))^2}{\lambda_n ||h_{2,1}||^2} = o_P(1), \tag{3.10}$$

eventually satisfy $\mathbb{P}_{\mathcal{H}}\gamma_n \in \mathbb{P}_{\mathcal{H}}\mathcal{S}_n$. The condition stated in (3.10) is equivalent to the condition that $||\mathbb{P}_{\mathcal{H}}\gamma_n||_\infty^2/\lambda_n \to 0$ as n tends to ∞. Let the term Condition I refer to the requirement that $\sqrt{n}\beta_{2,1}^*$ belong to $\mathcal{S}_n$ for all n. In addition, let Condition II refer to the requirement that for any sequence γ_n in $\mathbb{R}^p$ for which $||\mathbb{P}_{\mathcal{H}}\gamma_n||_\infty^2/\lambda_n \to 0$, it follows that $\mathbb{P}_{\mathcal{H}}\gamma_n \in \mathbb{P}_{\mathcal{H}}\mathcal{S}_n$ for all sufficiently large n. For any subset A of $\mathbb{R}^p$, let $\mathrm{diam}(A)$ denote $\sup_{a,b\in A}||a-b||$, the diameter of A. We have the following result.

Remark 3.1.1. Assume $\dim(\mathcal{H}) \geq 1$. Let $\mathcal{S}_n \triangleq \left\{\gamma \in \mathbb{R}^p \,:\, ||\Xi_{2,n}(\gamma - \sqrt{n}\beta_{2,1}^*)||_\infty^2 \leq \lambda_n\right\}$.

1. $\mathcal{S}_n$ satisfies Conditions I and II.

2. If $\mathcal{S}_n'$ is any other sequence of subsets of $\mathbb{R}^p$ satisfying Conditions I and II, then $\mathrm{diam}(\mathbb{P}_{\mathcal{H}}\mathcal{S}_n)/\mathrm{diam}(\mathbb{P}_{\mathcal{H}}\mathcal{S}_n') = O(1)$.

A proof of the preceding remark can be found in the supplement. The remark shows that the choice of $\mathcal{S}_n$ can only be tightened up to a constant factor without violating either Condition I or Condition II.

Suppose we want to construct a $1-\alpha$ confidence interval for $c^{\intercal}\beta_1^*$. By construction of $\mathcal{U}(c)$ and $\mathcal{L}(c)$ it follows that

$$P\left[c^{\intercal}\hat{\beta}_1 - \mathcal{U}(c)/\sqrt{n} \leq c^{\intercal}\beta_1^* \leq c^{\intercal}\hat{\beta}_1 - \mathcal{L}(c)/\sqrt{n}\right] \geq 1-\alpha.$$

We approximate the distribution of the $\mathcal{U}(c)$ and $\mathcal{L}(c)$ using the bootstrap. Namely, let $\hat{u}$ denote the $1-\alpha/2$ quantile of the bootstrap distribution of $\mathcal{U}(c)$, and let $\hat{l}$ denote the $\alpha/2$ quantile of the bootstrap distribution of $\mathcal{L}(c)$. Then $[c^{\intercal}\hat{\beta}_1 - \hat{u}/\sqrt{n}, c^{\intercal}\hat{\beta}_1 - \hat{l}/\sqrt{n}]$ is the ACI for $c^{\intercal}\beta_1^*$.

Next we show that the ACI is asymptotically valid. First define

1. $\Sigma_{t,\infty} \triangleq PB_t^{\intercal}B_t$ for $t = 1, 2$;
2. $g_2(B_2, Y_2; \beta_2^*) \triangleq B_2^{\intercal}(Y_2 - B_2\beta_2^*)$;
3. $g_1\left(B_1, Y_1, H_2; \beta_1^*, \beta_2^*\right) \triangleq B_1^{\intercal}\Big(Y_1 + H_{2,0}^{\intercal}\beta_{2,0}^* + \left[H_{2,1}^{\intercal}\beta_{2,1}^*\right]_+ - B_1\beta_1^*\Big)$;

We use the following assumptions.

(A1) The histories H_2, features B_1, and outcomes Y_t, satisfy the moment inequalities $P||H_2||^2\,||B_1||^2 < \infty$ and $PY_2^2||B_2||^2 < \infty$.

(A2) The matrices $\Sigma_{t,\infty}$ and $\mathrm{Cov}\left(g_1, g_2\right)$ are strictly positive definite.

(A3) The sequence λ_n tends to infinity and satisfies $\lambda_n = o(n)$.

(A4) For $\gamma \in \mathbb{R}^{\dim(\beta_{2,1}^*)}$, there exists $P_{n,\gamma}$ a sequence of contiguous alternatives converging to P in the sense that:

$$\int\left[\sqrt{n}\left(dP_{n,\gamma}^{1/2} - dP^{1/2}\right) - \frac{1}{2}g dP^{1/2}\right]^2 \to 0,$$

for some measurable function g for which if $\delta_2^{(n)} \triangleq \arg\min_{\delta} P_{n,\gamma}(Y_2 - Q_2(H_2, A_2; \delta))^2$, then $\delta_{2,1}^{(n)} \triangleq \beta_{2,1}^* + \gamma/\sqrt{n}$.

Assumptions (A1)-(A2) are quite mild, requiring only full rank design matrices and some moment conditions. Requirement (A3) regards a user-chosen tuning parameter and thus is always satisfied by appropriate choice of λ_n. Local alternatives provide a medium through which a glimpse of small sample behavior might be obtained, while retaining the mathematical convenience of large samples. Assumption (A4) facilitates a discussion of local alternatives without attempting to make the weakest possible assumptions (see van der Vaart and Wellner 1996).

The first result regards the population upper bound $\mathcal{U}(c)$.

Theorem 3.1.2 (Validity of population bounds)**.** *Assume (A1)-(A3) and fix* $c \in \mathbb{R}^{\dim(\beta_1^*)}$.

1. $c^\intercal \sqrt{n}(\hat{\beta}_1 - \beta_1^*) \rightsquigarrow c^\intercal \mathbb{W}_\infty + c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal H_{2,1}^\intercal \mathbb{V}_\infty 1_{H_{2,1}^\intercal \beta_{2,1}^* > 0} + c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal \left[H_{2,1}^\intercal \mathbb{V}_\infty \right]_+ 1_{H_{2,1}^\intercal \beta_{2,1}^* = 0}$.

2. *If for each n, the underlying generative distribution is* $P_{n,\gamma}$*, which satisfies (A4), then the limiting distribution of* $c^\intercal \sqrt{n}(\hat{\beta}_1 - \beta_1^*)$ *is given by:*

$$c^\intercal \mathbb{W}_\infty + c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal H_{2,1}^\intercal \mathbb{V}_\infty 1_{H_{2,1}^\intercal \beta_{2,1}^* > 0} \\ + c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal \left(\left[H_{2,1}^\intercal (\mathbb{V}_\infty + \gamma) \right]_+ - \left[H_{2,1}^\intercal \gamma \right]_+ \right) 1_{H_{2,1}^\intercal \beta_{2,1}^* = 0}. \quad (3.11)$$

3. *The limiting distribution of* $\mathcal{U}(c)$ *under both* P *and under* $P_{n,\gamma}$ *is equal to*

$$c^\intercal \mathbb{W}_\infty + c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal H_{2,1}^\intercal \mathbb{V}_\infty 1_{H_{2,1}^\intercal \beta_{2,1}^* > 0} \\ + \sup_{\gamma\prime \in \mathbb{R}^{\dim(\beta_{2,1}^*)}} c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal \left(\left[H_{2,1}^\intercal (\mathbb{V}_\infty + \gamma\prime) \right]_+ - \left[H_{2,1}^\intercal \gamma\prime \right]_+ \right) 1_{H_{2,1}^\intercal \beta_{2,1}^* = 0}, \quad (3.12)$$

where $\mathbb{W}_\infty$ *and* $\mathbb{V}_\infty$ *are jointly asymptotically normal with mean zero.*

See the supplementary material for the proof. Notice that limiting distributions of $c^\intercal \sqrt{n}(\hat{\beta}_1 - \beta_1^*)$ and $\mathcal{U}(c)$ (or equivalently $\mathcal{L}(c)$) are equal in the case $H_{2,1}^\intercal \beta_{2,1}^* \neq 0$ with probability one. That is, when there is a large treatment effect for almost all patients then the upper (or lower) bound is tight. However, when there is a non-null subset of patients for which there is no treatment effect, then the limiting distribution of the upper bound is stochastically larger then the limiting distribution of $c^{\intercal}\sqrt{n}(\hat{\beta}_1 - \beta_1^*)$. Thus, the ACI adapts to the setting in which all patients experience a treatment effect.

Because the distribution of (3.11) depends on the local alternative, γ, $\hat{\beta}_1$ is a nonregular estimator (van der Vaart and Wellner, 1996). One might hope to construct an estimator of the distribution of (3.11) and use this estimator to approximate the distribution of $c^{\intercal}\sqrt{n}(\hat{\beta}_1 - \beta_1^*)$. However a consistent estimator of the distribution of (3.11) does not exist because $P_{n,\gamma}$ is contiguous with respect to P (assumption A4). To see this, let $F_{\gamma}(u)$ be the distribution of (3.11) evaluated at a point, u. If a consistent estimator, say $\hat{F}_n(u)$, existed, that is $\hat{F}_n(u)$ converges in probability to $F_{\gamma}(u)$ under $P_{n,\gamma}$ then the contiguity implies that $\hat{F}_n(u)$ converges in probability to $F_{\gamma}(u)$ under P. This is a contradiction (at best $\hat{F}_n(u)$ converges in probability to $F_0(u)$ under P). Because we can not consistently estimate γ and we don't know the value of γ, the tightest, estimable upper bound on (3.11) is given by (3.12). As we shall next see, we are able to consistently estimate the distribution of (3.12).

In order to form confidence sets the bootstrap distributions of $\mathcal{U}(c)$ and $\mathcal{L}(c)$ are used. The next result regards the consistency of these bootstrap distributions. Let $\hat{\mathbb{P}}_n^{(b)}$ denote the bootstrap empirical measure, that is, $\hat{\mathbb{P}}_n^{(b)} \triangleq n^{-1}\sum_{i=1}^{n} M_{n,i}\delta_{\mathcal{T}_i}$ for $M_{n,1}, M_{n,2}, \ldots, M_{n,n} \sim \text{Multinomial}(n, (1/n, 1/n, \ldots, 1/n))$. We use the superscript (b) to denote that a functional has been replaced by its bootstrap analogue, so that if $\omega \triangleq f(\mathbb{P}_n)$ then $w^{(b)} \triangleq f(\hat{\mathbb{P}}_n^{(b)})$. Denote the space of bounded Lipschitz-1 functions on $\mathbb{R}^2$ by $BL_1(\mathbb{R}^2)$. Furthermore, let $\mathbb{E}_M$ and P_M denote the expectation and probability with respect to the bootstrap weights. The following results are proved in the supplemental material.

Theorem 3.1.3. *Assume (A1)-(A3) and fix* $c \in \mathbb{R}^{\dim(\beta_1^*)}$*. Then* $(\mathcal{U}(c), \mathcal{L}(c))$ *and* $(\mathcal{U}^{(b)}(c), \mathcal{L}^{(b)}(c))$ *converge to the same limiting distribution in probability. That is,*

$$\sup_{v \in BL_1(\mathbb{R}^2)} \left| \mathbb{E}v\left((\mathcal{U}(c), \mathcal{L}(c))\right) - \mathbb{E}_M v\left(\left(\mathcal{U}^{(b)}(c), \mathcal{L}^{(b)}(c)\right)\right) \right|$$

converges in probability to zero.

Corollary 3.1.4. *Assume (A1)-(A3) and fix* $c \in \mathbb{R}^{\dim(\beta_1^*)}$*. Let* $\hat{u}$ *denote the* $1 - \alpha/2$ *quantile of* $\mathcal{U}^{(b)}(c)$ *and* $\hat{l}$ *denote the* $\alpha/2$ *quantile of* $\mathcal{L}^{(b)}(c)$*. Then*

$$P_M\left(c^{\intercal}\hat{\beta}_1 - \hat{u}/\sqrt{n} \leq c^{\intercal}\beta_1^* \leq c^{\intercal}\hat{\beta}_1 - \hat{l}/\sqrt{n}\right) \geq 1 - \alpha + o_P(1).$$

Furthermore, if $P(H_{2,1}^{\intercal}\beta_{2,1}^* = 0) = 0$*, then the above inequality can be strengthened to equality.*

The preceding results show that the ACI can be use to construct valid confidence intervals regardless of the underlying parameters or generative model. Moreover, in settings where there is a treatment effect for almost every patient, the ACI delivers asymptotically exact coverage. See Section 4 for discussion of the choice of the tuning parameter λ_n.

3.2 Extending the ACI to many treatments

The two stage binary treatment setting which was addressed in the previous section provides the tools necessary to analyze data from many SMART trials including the ADHD study. Nonetheless, there is interest in analyzing data from multistage randomized trials where more than two treatments are available at each stage (Rush et al. 2003; Lieberman et al. 2005). In this section, we extend the ACI procedure for use with two stage trials with an arbitrary number of treatments available at each stage. The organization of this section parallels that of the previous section, however, the material is presented in a somewhat abbreviated fashion since much of the intuition has already been provided in earlier sections. In order to develop the results in this section, we require additional notation. Again, we observe trajectories $(X_1, A_1, Y_1, X_2, A_2, Y_2)$ drawn *i.i.d.* from some fixed and unknown distribution P. The treatment actions A_t take values in the set $\{1, \ldots, K_t\}$ for some fixed number of treatments K_t. In typical studies, K_t is no greater than five. We assume that the treatment action A_t is randomized with probabilities possibly depending on patient history, H_t ($H_t = \{X_1, A_1, \ldots, X_t\}$). We use the following linear model for the Q-function at time t:

$$Q_t(h_t, a_t; \beta_t) \triangleq \beta_{t,0}^{\intercal} h_{t,0} + \sum_{i=1}^{K_t} \beta_{t,i}^{\intercal} h_{t,1} 1_{a_t=i} \tag{3.13}$$

where as before $h_{t,0}$, $h_{t,1}$ are vectors of patient features constructed from the patient history, h_t and $\beta_t \triangleq (\beta_{t,0}^{\intercal}, \beta_{t,1}^{\intercal}, \ldots, \beta_{t,K_t}^{\intercal})^{\intercal}$. For identifiability, we assume that the vector of coefficients β_t satisfies a zero-sum constraint. That is, for each $j = 1, 2, \ldots, \dim(\beta_{t,1})$ the vector β_t satisfies $\sum_{i=1}^{K_t} \beta_{t,i,j} = 0$. The encoding of the treatments in this model is different than was used in the two stage binary treatment case. The reason for this difference is two-fold. The first reason is simply due to notation. Using a baseline constraint leads to much more complex expressions in this context. Second, setting $K_2 = 2$ and applying the results of this section illustrates how to deal with the two stage binary treatment setting when a contrast coding is employed. Thus we will have covered the two most commonly employed treatment codings. Note that according to this working model, if $h_{t,1}^{\intercal}\beta_{t,i} - \max_{j \neq i} h_{t,1}^{\intercal}\beta_{t,j} \approx 0$ for some $1 \leq i \leq K_t$, then at least two treatments are approximately optimal for a patient with history $H_{t,1} = h_{t,1}$. That is, there is *not* unique best treatment for a patient with history $H_{t,1} = h_{t,1}$. Conversely, if $\left|h_{t,1}^{\intercal}\beta_{t,i} - \max_{j\neq i} h_{t,1}^{\intercal}\beta_{t,j}\right| \gg 0$ for all $1 \leq i \leq K_t$, then *exactly* one treatment yields the best expected outcome for a patient with history $H_{t,1} = h_{t,1}$. As before, estimation of the optimal DTR is done using the Q-learning algorithm. The Q-learning algorithm proceeds as follows:

1. Regress Y_2 on H_2 and A_2 using (3.13) to obtain:

$$\hat{\beta}_2 \triangleq \arg\min_{\beta_2} \mathbb{P}_n(Y_2 - Q_2(H_2, A_2; \beta_2))^2,$$

   and subsequently the approximation $Q_2(h_2, a_2; \hat{\beta}_2)$ to the conditional mean $Q_2(h_2, a_2)$.

2. (a) Define the predicted future reward following the optimal policy as:

$$\begin{aligned} \tilde{Y}_1 &\triangleq Y_1 + \max_{a_2\in\{1,2,\ldots,K_2\}} Q_2(H_2, a_2; \hat{\beta}_2) && (3.14) \\ &= Y_1 + H_{2,0}^{\intercal}\hat{\beta}_{2,0} + \max_{1\leq i\leq K_2} H_{2,1}^{\intercal}\hat{\beta}_{2,i} && (3.15) \end{aligned}$$

   (b) Regress $\tilde{Y}_1$ on H_1 and A_1 using (3.13) to obtain $\hat{\beta}_1 \triangleq \arg\min_{\beta_1} \mathbb{P}_n(\tilde{Y}_1 - Q_1(H_1, A_1; \beta_1))^2$.

3. Define the estimated optimal DTR $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2)$ so that

$$\hat{\pi}_t(h_t) \triangleq \arg \max_{a_t\in\{1,2,\ldots,K_t\}} Q_t(h_t, a_t; \hat{\beta}_t).$$

As before, examination of the normal equations used to construct $\hat{\beta}_1$ combined with the definition of $\tilde{Y}_1$ show that $c^{\intercal}\sqrt{n}(\hat{\beta}_1 - \beta_1^*)$ can be decomposed as $c^{\intercal}\mathbb{W}_n + c^{\intercal}\Sigma_{n,1}^{-1}\mathbb{P}_n B_1^{\intercal}\mathbb{U}_n$, where the definitions of $\mathbb{W}_n$ and $\mathbb{U}_n$ have been generalized to

$$\begin{aligned}\mathbb{W}_n &= \Sigma_{1,n}^{-1}\sqrt{n}\mathbb{P}_n B_1^{\intercal}\left\{\left(Y_1 + H_{2,0}^{\intercal}\beta_{2,0}^* + \left[\max_{1\le i\le K_2} H_{2,1}^{\intercal}\beta_{2,i}^*\right] - B_1\beta_1^*\right) + H_{2,0}^{\intercal}(\hat{\beta}_{2,0} - \beta_{2,0}^*)\right\},\\ \mathbb{U}_n &= \sqrt{n}\left(\left[\max_{1\le i\le K_2} H_{2,1}^{\intercal}\hat{\beta}_{2,i}\right] - \left[\max_{1\le i\le K_2} H_{2,1}^{\intercal}\beta_{2,i}^*\right]\right).\end{aligned}$$

The non-regularity of the limiting distribution of $c^{\intercal}\sqrt{n}(\hat{\beta}_1 - \beta_1^*)$ is apparent by noting the non-differentiable max operator in the definition of $\mathbb{U}_n$. Define

$$\mathcal{A}_2^*(h_{2,1}) \triangleq \left\{\arg\max_{1\le i\le K_2} h_{2,1}^{\intercal}\beta_{2,i}^*\right\}$$

to be the set of equally optimal treatments for a patient with history $H_{2,1} = h_{2,1}$. Thus, $\mathcal{A}_2^*(h_{2,1})$ is a singleton when there is exactly one unique best treatment for a patient with history $h_{2,1}$. Define $\mathbb{V}_{n,i} \triangleq \sqrt{n}(\hat{\beta}_{2,i} - \beta_{2,i}^*)$ for $i = 1, 2, \ldots, K_2$. Let $\Xi_{2,n}'$ denote the inverse of the matrix square-root of the plug-in estimator of the asymptotic covariance of $(\mathbb{V}_{n,1}^{\intercal}, \ldots, \mathbb{V}_{n,K_2}^{\intercal})^{\intercal}$ and let $\beta_{2,:}^*$ denote the vector $(\beta_{2,1}^{*\intercal}, \ldots, \beta_{2,n}^{*\intercal})^{\intercal}$. The upper bound $\mathcal{U}(c)$ used to construct the ACI in the many treatment case is given by

$$\begin{aligned}& c^{\intercal}\mathbb{W}_n + c^{\intercal}\Sigma_{1,n}^{-1}\mathbb{P}_n B_1^{\intercal}\mathbb{U}_n 1_{\#\hat{\mathcal{A}}_2(H_{2,1})=1} + \\ & + \sup_{\gamma\in\mathcal{S}_n} c^{\intercal}\Sigma_{1,n}^{-1}\mathbb{P}_n B_1^{\intercal}\left(\max_{i\in\tilde{\mathcal{A}}_2(H_{2,1})} H_{2,1}^{\intercal}\left(\mathbb{V}_{n,i} + \gamma_i\right) - \max_{i\in\tilde{\mathcal{A}}_2(H_{2,1})} H_{2,1}^{\intercal}\gamma_i\right) 1_{\#\hat{\mathcal{A}}_2(H_{2,1})>1}, \quad (3.16)\end{aligned}$$

where $\gamma = (\gamma_1^{\intercal}, \gamma_2^{\intercal}, \ldots, \gamma_{K_2}^{\intercal})^{\intercal}$, $\mathcal{S}_n \triangleq \left\{\gamma \in \mathbb{R}^{\dim(\beta_{2,1}^*)} : ||\Xi_{2,n}(\gamma - \sqrt{n}\beta_{2,:}^*)||_\infty^2 \le \lambda_n\right\}$, and $\hat{\mathcal{A}}_2(h_{2,1})$ is an estimator of $\mathcal{A}_2^*(h_{2,1})$ and $\tilde{\mathcal{A}}_2(h_{2,1}) \triangleq \hat{\mathcal{A}}_2(h_{2,1}) \cup \mathcal{A}_2^*(h_{2,1})$. The estimator $\hat{\mathcal{A}}_2(h_{2,1})$ of $\mathcal{A}_2^*(h_{2,1})$ is based on a series of hypothesis tests using the test statistics:

$$T_{n,i}(h_{2,1}) \triangleq \frac{n\left(h_{2,1}^{\intercal}\hat{\beta}_{2,i} - \max_{j\ne i} h_{2,1}^{\intercal}\hat{\beta}_{2,j}\right)^2}{h_{2,1}^{\intercal}\hat{\zeta}_i h_{2,1}},$$

where $\hat{\zeta}_i$ is the usual plug-in estimator of $n\, Cov(\hat{\beta}_{2,i} - \hat{\beta}_{2,j})$ for $j = \arg\max_{j \neq i} h_{2,1}^{\intercal}\hat{\beta}_{2,j}$, assuming the index j to be fixed a priori (see Hsu 1996). Notice that $\min_i T_{n,i}(h_{2,1})$ should be large if treatment i is the uniquely optimal treatment for a patient with history $H_{2,1} = h_{2,1}$. On the other hand, $T_{n,i}(h_{2,1})$ should be small if treatment i is the optimal treatment for a patient with history $h_{2,1}$ *and* there is more than one best treatment. A natural estimator of $\mathcal{A}_2^*(h_{2,1})$ is

$$\hat{\mathcal{A}}_2(h_{2,1}) = \begin{cases} \{i : T_{n,i}(h_{2,1}) \leq \lambda_n\} & \text{if } \min_i T_{n,i}(h_{2,1}) \leq \lambda_n \\ \arg\max_{1 \leq i \leq K_2} h_{2,1}^{\intercal}\hat{\beta}_{2,i} & \text{if } \min_i T_{n,i}(h_{2,1}) > \lambda_n. \end{cases}$$

The test statistic, $T_{n,i}(h_{2,1})$, is taken from the “multiple comparisons with the best” literature (see Hsu 1996 and references therein). Let Δ denote the symmetric set difference. It can be shown that when λ_n diverges to infinity and satisfies $\lambda_n = o(n)$, then $\hat{\mathcal{A}}_2(h_{2,1})$ is a consistent estimator of $\mathcal{A}_2^*(h_{2,1})$ in the sense that $\#\left[\mathcal{A}_2^*(h_{2,1})\,\Delta\,\hat{\mathcal{A}}_2(h_{2,1})\right]$ converges to zero in probability.

The intuition behind the upper bound $\mathcal{U}(c)$ is the same as in the binary treatment case. Namely, we partition the data based into two sets (i) patients for which there is exactly one best treatment, and (ii) patients for which there is more than one best treatment. Different approximations are used on each partition to ensure asymptotic consistency and a bound that controls the influence of errors in the partitioning is added to ensure good small sample performance. The bootstrap distributions of the upper bound $\mathcal{U}(c)$ and lower bound $\mathcal{L}(c)$ (which is formed by replacing the sup with an inf in the definition of $\mathcal{U}(c)$) are used to construct a confidence set.

The theoretical results presented for the binary treatment ACI, including those regarding the bootstrap, hold in the many treatment case as well. While there is no qualitative change in the required assumptions, they must however be generalized to accommodate an arbitrary number of treatments. The generalized assumptions along with statements of the theorems in the many treatment case can be found in the supplementary material.

3.3 Empirical Study

In this section we contrast different choices of the potentially important tuning parameter λ_n and we provide an empirical evaluation of the ACI. Fourteen generative models are used in these evaluations; the first seven of these come from (Chakraborty et al. 2009). Each of these seven generative models has two stages of treatment and two treatments at each stage. The second seven generative models are similar but have three treatments at stage two; a complete description of these models can be found in the supplemental material. Generically, each of the first seven models can be described as follows:

- $X_i \in \{-1, 1\}$, $A_i \in \{-1, 1\}$ for $i \in \{1, 2\}$
- $P(A_1 = 1) = P(A_1 = -1) = 0.5$, $P(A_2 = 1) = P(A_2 = -1) = 0.5$
- $X_1 \sim \text{Bernoulli}(0.5)$, $X_2 | X_1, A_1 \sim \text{Bernoulli}(\text{expit}(\delta_1 X_1 + \delta_2 A_1))$
- $Y_1 \triangleq 0$,
  $Y_2 = \gamma_1 + \gamma_2 X_1 + \gamma_3 A_1 + \gamma_4 X_1 A_1 + \gamma_5 A_2 + \gamma_6 X_2 A_2 + \gamma_7 A_1 A_2 + \epsilon$, $\epsilon \sim N(0, 1)$

where $\text{expit}(x) = \mathrm{e}^x/(1+\mathrm{e}^x)$. This class is parameterized by nine values $\gamma_1, \gamma_2, ..., \gamma_7, \delta_1, \delta_2$. The analysis model uses patient feature vectors defined by:

$$\begin{aligned} H_{2,0} &= (1, X_1, A_1, X_1 A_1, X_2)^\intercal \\ H_{2,1} &= (1, X_2, A_1)^\intercal \end{aligned}$$

$$\begin{aligned} H_{1,0} &= (1, X_1)^\intercal \\ H_{1,1} &= (1, X_1)^\intercal. \end{aligned}$$

Our working models are given by $Q_2(H_2, A_2; \beta_2) \triangleq H_{2,0}^\intercal \beta_{2,0} + H_{2,1}^\intercal \beta_{2,1} A_2$ and $Q_1(H_1, A_1; \beta_1) \triangleq H_{1,0}^\intercal \beta_{1,0} + H_{1,1}^\intercal \beta_{1,1} A_1$. We use contrast encoding for A_1 and A_2 to allow for a comparison with Chakraborty et al. (2009).

The form of this class of generative models is useful as it allows us to influence the degree of non-regularity present in our example problems through the choice of the γ_i and δ_i, and in turn evaluate performance in these different scenarios. Recall that in Q-learning, non-regularity occurs when more than one stage-two treatment produces nearly the same optimal expected reward for a set of patient histories that occur with positive probability. In the model class above, this occurs if the model generates histories for which $\gamma_5 A_2 + \gamma_6 X_2 A_2 + \gamma_7 A_1 A_2 \approx 0$, i.e., if it generates histories for which Q_2 depends weakly or not at all on A_2. By manipulating the values of γ_i and δ_i, we can control i) the probability of generating a patient history such that $\gamma_5 A_2 + \gamma_6 X_2 A_2 + \gamma_7 A_1 A_2 = 0$, and ii) the standardized effect size $E[(\gamma_5 + \gamma_6 X_2 + \gamma_7 A_1)/\sqrt{\mathrm{Var}(\gamma_5 + \gamma_6 X_2 + \gamma_7 A_1)}]$. Each of these quantities, denoted by p and ϕ, respectively, can be thought of as measures of problem non-regularity. Six different generative models are provided by Chakraborty et al. (2009), and are described by them as "non-regular", "near-non-regular", and "regular". We have added one additional model to include the case of a strongly regular setting.

3.3.1 The choice of λ_n

We measure and compare the performance of four choices of the tuning parameter λ_n in terms of estimated coverage and average interval diameter. The intervals are constructed for intercept and the coefficient of the treatment indicator in the first

Example	γ	δ	Type		
1	$(0,0,0,0,0,0,0)^\intercal$	$(0.5,0.5)^\intercal$	non-regular	$p=1$	$\phi=0/0$
2	$(0,0,0,0,0.01,0,0)^\intercal$	$(0.5,0.5)^\intercal$	near-non-regular	$p=0$	$\phi=\infty$
3	$(0,0,-0.5,0,0.5,0,0.5)^\intercal$	$(0.5,0.5)^\intercal$	non-regular	$p=1/2$	$\phi=1.0$
4	$(0,0,-0.5,0,0.5,0,0.49)^\intercal$	$(0.5,0.5)^\intercal$	near-non-regular	$p=0$	$\phi=1.0204$
5	$(0,0,-0.5,0,1.0,0.5,0.5)^\intercal$	$(1.0,0.0)^\intercal$	non-regular	$p=1/4$	$\phi=1.4142$
6	$(0,0,-0.5,0,0.25,0.5,0.5)^\intercal$	$(0.1,0.1)^\intercal$	regular	$p=0$	$\phi=0.3451$
7	$(0,0,-0.25,0,0.75,0.5,0.5)$	$(0.1,0.1)$	regular	$p=0$	$\phi=1.035$

Table 3.1: Parameters indexing the example models.

stage Q-function in the fourteen generative models. We use a training set size of $n = 150$ in order to mimic the sample size of the ADHD study ($n = 138$). The online supplement contains a number of additional examples and sample sizes all displaying similar trends as presented here. For the sequence λ_n we consider the following settings: $\lambda_n = \sqrt{\log\log\, n}$, $\log\log\, n$, $log\, n$, n.

The intuition behind these settings is as follows. The supremum (infimum) used in the ACI can be thought of controlling the influence of committing a Type II error in the test of $\mathcal{N}_0(h_{2,1}) : h_{2,1}^\intercal\beta_{2,1}^* = 0$. On the other hand, the Type I error is controlled by the choice of λ_n. Recall that we reject the hypothesis $\mathcal{N}_0(h_{2,1})$ when $T_n(h_{2,1}) > \lambda_n$. Thus, it is of interest to examine the (uniform) behaviour of $T_n(h_{2,1})/\lambda_n$ across the set of $h_{2,1}$ for which $\mathcal{N}_0(h_{2,1})$ is true. Since the test statistic T_n is scale invariant (e.g. for any $\alpha > 0$ we have $T_n(\alpha h_{2,1}) = T_n(h_{2,1})$) is suffices to restrict our attention to unit vectors $h_{2,1}$ satisfying $\mathcal{N}_0(h_{2,1})$. We let $\mathcal{W} \triangleq \left\{h_{2,1} \in \mathbb{R}^{\dim(\beta_{2,1}^*)} \,:\, h_{2,1}^\intercal\beta_{2,1}^* = 0,\, ||h_{2,1}|| = 1\right\}$ denote these vectors of interest. Provided that λ_n tends to ∞ it follows that

$$\sup_{h\in\mathcal{W}} T_n(h)/\lambda_n \to 0$$

in probability. Furthermore, if λ_n grows faster than $\log\log\, n$ then the above convergence can be strengthened from in probability to almost surely using the law of the iterated logarithm (see Csorgo and Rosalsky 2003). However, consistency of the ACI also requires that $\lambda_n = o(n)$. Thus, $\lambda_n = n$ represents rate that is too fast for consistency to hold; $\lambda_n = \log n$ is fast enough for strong (almost sure) control of the Type I error; $\lambda_n = \log\, \log n$ represents a rate that is at the boundary between almost sure and in convergence in probability; $\lambda_n = \sqrt{\log \log\, n}$ represents a rate that only ensures convergence in probability.

Tables (3.4) and (3.5) show the estimated coverage and interval diameter of the ACI across the four parameter settings for the first seven generative models. The results appear stable across all choices of λ_n. However, the ACI seems excessively conservative when λ_n is allowed to grow faster than log log n. Both in the simulation studies below as well as in the data analysis, we use $\lambda_n = \log\, \log\, n$.

3.3.2 An Evaluation of the ACI

We compare the empirical performance of the ACI with the centered percentile bootstrap (CPB), the soft-thresholding (ST) method of Chakraborty et al. (2009), and simple plug-in pretesting estimator (PPE). The hard-thresholding of Moodie and Richardson (2007) is similar in theory and performance to the soft-thresholding approach; furthermore in orthogonal settings the lasso type penalization of Song et al. (2010) is equivalent to soft-thresholding, and so, Chakraborty's method is used to represent these alternate approaches. The performance of each method is measured in terms of estimated coverage and interval diameter. We shall see that the ACI is conservative when there is no stage 2 treatment effect for all feature patterns; this is not unexpected since the ACI is based on the use of the upper/lower bound. Despite the use of the bounds, ACI routinely delivers close to the nominal coverage and possesses competitive diameters. Competing methods fail to attain nominal coverage on many of the examples.

C.I.s for $\beta_{1,1,1}$	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
$\lambda_n =$	NR	NNR	NR	NNR	NR	R	R
$\sqrt{\log(\log(n))}$	0.9871	0.9869	0.9628	0.9645	0.9545	0.9515	0.9533
$\log(\log(n))$	0.9895	0.9891	0.9646	0.9662	0.9560	0.9544	0.9545
$\log(n)$	0.9951	0.9953	0.9700	0.9728	0.9628	0.9663	0.9615
n	0.9967	0.9968	0.9714	0.9738	0.9652	0.9737	0.9684

C.I.s for $\beta_{1,0,1}$	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
$\lambda_n =$	NR	NNR	NR	NNR	NR	R	R
$\sqrt{\log(\log(n))}$	0.9622	0.9692	0.9631	0.9657	0.9575	0.9608	0.9559
$\log(\log(n))$	0.9652	0.9709	0.9651	0.9673	0.9592	0.9624	0.9566
$\log(n)$	0.9726	0.9780	0.9706	0.9727	0.9631	0.9730	0.9648
n	0.9739	0.9787	0.9719	0.9743	0.9660	0.9799	0.9719

Table 3.2: Monte Carlo estimates of coverage probabilities for the ACI methods at the 95% nominal level. Here, $\beta_{1,1,1}$ denotes the main effect of treatment and $\beta_{1,0,1}$ denotes the intercept. Estimates are constructed using 10000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Examples are designated NR = non-regular, NNR = near-non-regular, R = regular.

C.I.s for $\beta_{1,1,1}$	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
$\lambda_n =$	NR	NNR	NR	NNR	NR	R	R
$\sqrt{\log(\log(n))}$	0.4890	0.4890	0.4795	0.4796	0.4816	0.4681	0.4711
$\log(\log(n))$	0.5008	0.5007	0.4857	0.4857	0.4847	0.4717	0.4735
$\log(n)$	0.5560	0.5560	0.5156	0.5155	0.5009	0.4914	0.4883
n	0.5809	0.5808	0.5311	0.5311	0.5113	0.5099	0.5081

C.I.s for $\beta_{1,0,1}$	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
$\lambda_n =$	NR	NNR	NR	NNR	NR	R	R
$\sqrt{\log(\log(n))}$	0.5056	0.5057	0.4794	0.4795	0.4815	0.4867	0.4713
$\log(\log(n))$	0.5177	0.5176	0.4856	0.4856	0.4846	0.4907	0.4739
$\log(n)$	0.5732	0.5731	0.5159	0.5159	0.5011	0.5135	0.4909
n	0.5942	0.5942	0.5340	0.5343	0.5130	0.5393	0.5172

Table 3.3: Monte Carlo estimates of mean width of the ACI method at the 95% nominal level. Here, $\beta_{1,1,1}$ denotes the main effect of treatment and $\beta_{1,0,1}$ denotes the intercept. Estimates are constructed using 10000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Models have two treatments at each of two stages. Examples are designated NR = non-regular, NNR = near-non-regular, R = regular.

We now briefly describe the competing methods. A natural first method to try is the bootstrap; thus it serves as a useful baseline for comparison. As discussed the ST method works by shrinking the fitted regression $\hat{\beta}_1$ in the hopes of mitigating bias induced by non-regularity. In particular, for the working models we consider in this section; the ST estimators are:

$$\hat{\beta}_1^{ST} \triangleq \arg\min_{\beta_1} \mathbb{P}_n(\tilde{Y}_1^{ST} - B_1\beta_1)^2 \tag{3.17}$$

$$\tilde{Y}_1^{ST} \triangleq Y_1 + H_{2,0}^{\intercal}\hat{\beta}_{2,0} + |H_{2,1}^{\intercal}\hat{\beta}_{2,1}|\left(1 - 3\frac{H_{2,1}^{\intercal}\Sigma_{2,n}^{(2,2)}H_{2,1}}{n|H_{2,1}^{\intercal}\hat{\beta}_{2,1}|}\right)_+ . \tag{3.18}$$

In the above display, $\hat{\beta}_2$ and $\Sigma_{2,n}^{(2,2)}$ are as described in previous sections. The constant 3 appearing in the ST method is motivated by an empirical Bayes interpretation of the thresholding (see work by Chakraborty et al. (2009) for more details). The form of the ST method shows that the modified predicted future reward following the optimal policy is shrunk most heavily when $h_{2,1}^{\intercal}\hat{\beta}_{2,1}$ is small. Which is to say, shrinkage occurs when there is little evidence that one treatment differs significantly from another for a patient with history $H_{2,1} = h_{2,1}$. The ST method is only developed for binary treatment.

The PPE confidence interval, in the two-stage binary treatment case, is formed by bootstrapping

$$c^{\intercal}\mathbb{W}_n + c^{\intercal}\Sigma_{1,n}^{-1}\mathbb{P}_nB_1^{\intercal}\mathbb{U}_n 1_{T_n(H_{2,1})>\lambda_n} + c^{\intercal}\Sigma_{1,n}^{-1}\mathbb{P}_nB_1^{\intercal}\left[H_{2,1}^{\intercal}\mathbb{V}_n\right]_+ 1_{T_n(H_{2,1})\leq\lambda_n}. \tag{3.19}$$

This approach is natural as it partitions the data using a pretest and then uses a different estimator on each partition. A similar idea was employed by Chatterjee and Lahiri (2009) in their treatment of the Lasso. However, this approach is consistent under fixed but not local alternatives (see the supplemental material for additional details; see also Leeb and Potscher 2005). As we will see below, this leads to rather poor small sample performance. The primary reason for including this method is to motivate the importance of local alternatives and the utility of the supremum (infimum) in the construction of the ACI.

We first provide confidence intervals for the coefficient of A_1 (the treatment variable), $\beta^*_{1,1,1}$. Note that given the working models and generative models defined by the parameter settings in Table 3.24, we can determine the exact value of any parameter $c^\intercal\beta^*_1$ of interest. The supplementary material contains confidence intervals for the treatment effect when $X_1 = 1$ (e.g. $\beta^*_{1,1,1} + \beta^*_{1,1,2}$). In addition, it contains estimated coverage probabilities and interval diameters for a range of sample sizes and a number of additional generative models, including those with multiple stages of treatment.

Table 3.4 shows the estimated coverage for the coefficient of A_1, $\beta^*_{1,1,1}$. This simulation uses a sample size of 150, a total of 1000 Monte Carlo replications and 1000 bootstrap samples. Target coverage is .95. The CPB and PPE methods fare the worst in terms of coverage, each falling significantly below nominal coverage on ten of the fourteen examples respectively. The ST method fails to cover in the most regular, the seventh, example. The reason for this under performance is that the ST method tends to over-shrink when there are large treatment effects. Recall that the ST method has not been developed for the setting in which there are more than two treatments at the second stage. The ACI is the only method to deliver nominal coverage on all fourteen examples. The ACI is conservative on examples one and two. The average interval diameters are shown in Table 3.5. The ACI is the most conservative as is to be expected given that it is based on upper and lower bounds. However, the width is non-trivial and is actually the smallest in several of the examples if one only considers methods that attain nominal coverage.

The coefficient of A_1 is perhaps most relevant from a clinical perspective. However, from a methodological point of view, other contrasts can be illuminating. Table 3.6 shows the estimated coverage for intercept using the same generative models described in the preceding paragraph. The coverage of competing methods is quite poor collectively attaining nominal coverage on two examples. Particularly disturbing is that the ST method falls more than 30% below nominal levels. In contrast, the ACI delivers nominal coverage on every example. Table 3.7 shows the average interval widths; the ACI is the widest but non-trivial.

3.4 Analysis of the ADHD study

In this section we illustrate the use of the ACI on data from the Adaptive Interventions for Children with ADHD study (Pehlam et al. 2008). The ADHD data we use here consists of $n = 138$ trajectories. These $n = 138$ trajectories form a subset the original $N = 155$ observations. This subset was formed by removing the $N - n = 17$ patients that were either never randomized to an initial (first stage) treatment (14 patients), or had massive item missingness (3 patients). A description of each of the variables is described in Table (3.8). Notice that the outcomes Y_1 and Y_2 satisfy $Y_1 + Y_2 \equiv R$, where R is the teacher reported TIRS5 score at the last week of the study (week 32).

The Q-learning algorithm detailed in earlier sections of this paper describes how to estimate an optimal DTR. However, an estimation procedure alone is insufficient for model building. Tools for model assessment and criticism are also necessary. One of the chief advantages of Q-learning is that it consists of a series of linear regression models. Consequently, exploratory data analysis, model assessment and criticism (e.g. residual analysis) can be performed for each of the intermediate linear models

Two Treatments at Stage 2	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.934*	0.935*	0.930*	0.933*	0.938	0.928*	0.939
PPE	0.931*	0.940	0.938	0.940	0.946	0.912*	0.931*
ST	0.948	0.945	0.938	0.942	0.952	0.943	0.919*
ACI	0.992	0.992	0.968	0.972	0.957	0.955	0.950

Three Treatments at Stage 2	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.933*	0.937	0.915*	0.921*	0.931*	0.907*	0.940
PPE	0.931*	0.932*	0.927*	0.919*	0.932*	0.883*	0.919*
ACI	0.996	0.996	0.968	0.969	0.959	0.968	0.959

Table 3.4: Monte Carlo estimates of coverage probabilities of confidence intervals for the coefficient of the treatment variable, $\beta^*_{1,1,1}$ at the 95% nominal level. Estimates are constructed using 1000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. There is no ST method when there are three treatments at Stage 2. Examples are designated NR = non-regular, NNR = near-non-regular, R = regular.

Two Treatments at Stage 2	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.385*	0.385*	0.430*	0.430*	0.457	0.436*	0.451
PPE	0.365*	0.366	0.419	0.419	0.452	0.418*	0.452*
ST	0.339	0.339	0.426	0.427	0.469	0.436	0.480*
ACI	0.502	0.502	0.488	0.488	0.487	0.475	0.477

Three Treatments at Stage 2	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.446*	0.446	0.518*	0.518*	0.567*	0.518*	0.557
PPE	0.416*	0.416*	0.501*	0.501*	0.557*	0.487*	0.549*
ACI	0.655	0.655	0.625	0.625	0.621	0.630	0.616

Table 3.5: Monte Carlo estimates of mean width of confidence intervals for the coefficient of the treatment variable, $\beta^*_{1,1,1}$ at the 95% nominal level. Estimates are constructed using 1000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Models have two treatments at each of two stages. Widths with corresponding coverage significantly below nominal are marked with $*$. There is no ST method when there are three treatments at Stage 2. Examples are designated NR = non-regular, NNR = near-non-regular, R = regular.

Two Treatments at Stage 2	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.892*	0.908*	0.924*	0.925*	0.940	0.930*	0.936
PPE	0.926*	0.930*	0.933*	0.934*	0.934*	0.907*	0.928*
ST	0.935*	0.930*	0.889*	0.878*	0.891*	0.620*	0.687*
ACI	0.956	0.964	0.954	0.955	0.950	0.957	0.948

Table 3.6: Monte Carlo estimates of coverage probabilities of confidence intervals for the coefficient of the intercept, $\beta^*_{1,0,1}$ at the 95% nominal level. Estimates are constructed using 1000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Examples are designated NR = non-regular, NNR = near-non-regular, R = regular.

Two Treatments at Stage 2	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.404*	0.404*	0.430*	0.429*	0.457	0.449*	0.450
PPE	0.376*	0.376*	0.418*	0.418*	0.451*	0.448*	0.453*
ST	0.344*	0.344*	0.427*	0.427*	0.466*	0.469*	0.474*
ACI	0.518	0.518	0.487	0.487	0.486	0.494	0.476

Table 3.7: Monte Carlo estimates of mean width of confidence intervals for the coefficient of the intercept, $\beta^*_{1,0,1}$ at the 95% nominal level. Estimates are constructed using 1000 datasets of size 150 drawn from each model, and 1000 bootstraps drawn from each dataset. Models have two treatments at each of two stages. Widths with corresponding coverage significantly below nominal are marked with $*$. Examples are designated NR = non-regular, NNR = near-non-regular, R = regular.

$X_{1,1} \in [0, 3]$	:	baseline teacher reported mean ADHD symptom score. Measured at the end of the school year preceding the study.
$X_{1,2} \in \{0, 1\}$	:	indicator of a diagnosis of ODD (oppositional defiant disorder) at baseline, coded so that $X_{1,3} = 0$ corresponds to no such diagnosis.
$X_{1,3} \in \{0, 1\}$	:	indicator of a patient's prior exposure to ADHD medication, coded so that $X_{1,2} = 0$ corresponds to no prior exposure.
$A_1 \in \{-1, 1\}$	:	initial treatment, coded so that $A_1 = -1$ corresponds to medication while $A_1 = 1$ corresponds to behavioral modification therapy.
$T \in \{6, 7, \ldots 32\}$	:	right censored time in weeks until patient is re-randomized.
$Y_1 \triangleq R1_{T \geq 32}$	:	first stage response (see definition of R below).
$X_{2,1} \in \{0, 1\}$	:	indicator of patient's adherence to their initial treatment. Adherence is coded so that a value of $X_{2,1} = 0$ corresponds to low adherence (taking less than 100% of prescribed medication or attending less than 75% of therapy sessions) while a value of $X_{2,1} = 1$ corresponds to high adherence.
$X_{2,2} \in \{1, 8\}$	:	month of non-response.
$A_2 \in \{-1, 1\}$	:	second stage treatment, coded so that $A_2 = -1$ corresponds to augmenting the initial treatment with the treatment *not* received initially, and $A_2 = 1$ corresponds to enhancing (increasing the dosage of) the initial treatment.
$R \in \{1, 2, \ldots, 5\}$	:	teacher reported Teacher Impairment Rating Scale (TIRS5) item score 32 weeks after initial randomization to treatment. The TIRS5 is coded so that higher values correspond to better clinical outcomes.
$Y_2 \triangleq R1_{T<32}$	:	second stage outcome.

Table 3.8: Components of a single trajectory in the ADHD study.

by application of standard methods. Of course this approach is imperfect in that examining the individual linear models does not ensure that the combined model is appropriate.

The first step in estimating an optimal DTR from the ADHD study is to build a regression model for the second stage. The second stage regression model is built only using patients that were re-randomized during the 32 week study. Of the $n = 138$ patients, 79 of them were re-randomized before the study conclusion. The feature vectors at the second stage are $H_{2,0} \triangleq (1, X_{1,1}, X_{1,2}, X_{2,2}, X_{1,3}, X_{2,1}, A_1)^\intercal$ and $H_{2,1} \triangleq (1, X_{2,1}, A_1)^\intercal$. Thus, the Q-function $Q_2(H_2, A_2; \beta_2) \triangleq H_{2,0}^\intercal \beta_{2,0} + H_{2,1}^\intercal \beta_{2,1} A_2$

contains an interaction term between the second stage action A_2 and a patient's initial treatment A_1, an interaction between A_2 and adherence to their initial medication $X_{2,1}$, a main effect for A_2, and main effects for all the other terms. A table of the second stage least squares coefficients along with interval estimates are given in Table (3.9). Examination of the residuals (not shown here) shows no obvious signs of model misspecification. In short, the linear model described above seems to fit the data reasonably well.

Term	Coeff.	Estimate	Lower (5%)	Upper (95%)
Intercept	$\beta_{2,0,1}$	1.36	0.48	2.26
Baseline symptoms	$\beta_{2,0,2}$	0.94	0.48	1.39
ODD diagnosis	$\beta_{2,0,3}$	0.92	0.46	1.41
Month of non-response	$\beta_{2,0,4}$	0.02	-0.20	0.20
Prior Medication	$\beta_{2,0,5}$	-0.27	-0.77	0.21
Adherence	$\beta_{2,0,6}$	0.17	-0.28	0.66
First stage txt	$\beta_{2,0,7}$	0.03	-0.18	0.23
Second stage txt	$\beta_{2,1,1}$	-0.72	-1.13	-0.35
Second stage txt : Adherence	$\beta_{2,1,2}$	0.97	0.48	1.52
Second stage txt : First stage txt	$\beta_{2,1,3}$	0.05	-0.17	0.27

Table 3.9: Least squares coefficients and 90% interval estimates for second stage regression.

Recall that the response for this first stage regression model is the predicted future outcome $\tilde{Y}_1 \triangleq Y_1 + \max_{a_2 \in \{-1,1\}} Q_2(H_2, a_2; \hat{\beta}_2)$. Since the predictors used in the first stage must predate the assignment of first treatment, the available predictors in Table (3.8) are baseline ADHD symptoms $X_{1,1}$, diagnosis of ODD at baseline $X_{1,2}$, indicator of a patient's prior exposure to ADHD medication $X_{1,3}$, and first stage treatment A_1. The feature vectors for the second stage are $H_{1,0} \triangleq (1, X_{1,1}, X_{1,2}, X_{1,3})$ and $H_{1,1} \triangleq (1, X_{1,3})$, so that the first stage Q-function $Q_1(H_1, A_1; \beta_1) \triangleq H_{1,0}^{\intercal}\beta_{1,0} + H_{1,1}^{\intercal}\beta_{1,1}A_1$ contains an interaction term between the first stage action A_1 and a patient's prior exposure to ADHD medication $X_{1,3}$, a main effect for A_1, and main effects for all other covariates. The first stage regression coefficients are estimated using least squares $\hat{\beta}_1 \triangleq \arg\min_{\beta_1} \mathbb{P}_n(\tilde{Y}_1 - Q_1(H_1, A_1; \beta_1))^2$. A table of least squares coefficients along with interval estimates formed using the ACI are given in Table (3.10). Plots of the residuals for this model (not shown here) show no obvious signs of model misspecification. This is to say that the linear model seems to provide a reasonable approximation to the Q-function in the first stage.

Term	Coeff.	Estimate	Lower (5%)	Upper (95%)
Intercept	$\beta_{1,0,1}$	2.61	2.07	3.05
Baseline symptoms	$\beta_{1,0,2}$	0.73	0.46	1.02
ODD diagnosis	$\beta_{1,0,3}$	0.75	0.38	1.10
Prior med. exposure	$\beta_{1,0,4}$	-0.37	-0.79	0.01
Initial txt	$\beta_{1,1,1}$	0.17	-0.05	0.38
Initial txt : Prior med. exposure	$\beta_{1,1,2}$	-0.32	-0.61	-0.06

Table 3.10: Least squares coefficients and 90% ACI interval estimates for first stage regression.

Having fit both the first and second stage regressions, we now construct an estimate of the optimal DTR. Recall that for any $H_t = h_t,\ t = 1, 2$ the estimated optimal DTR $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2)$ satisfies $\hat{\pi}_t(h_t) \in \arg\max_{a_t} Q(h_t, a_t; \hat{\beta}_t)$. The coefficients in Table (3.9) and the form of the second stage Q-function reveal that the second stage decision rule $\hat{\pi}_2$ is quite simple. In particular, $\hat{\pi}_2$ prescribes treatment enhancement to patients with high adherence to their initial medication and it prescribes treatment augmentation to patients with low adherence to their initial medication. The first stage decision rule $\hat{\pi}_1$ is equally simplistic. The coefficients in Table (3.10) show that the first stage decision rule, $\hat{\pi}_1$ prescribes medication to patients that have had prior exposure to medication, and behavioral modification to patients that have not had any such prior exposure.

The prescriptions given by the estimated optimal DTR $\hat{\pi}$ are excessively decisive.

That is, they recommend one and only one treatment regardless of the amount of evidence in the data to support that the recommended treatment is in fact optimal. When there is insufficient evidence to recommend a single treatment as best for a given patient history, it is preferred to leave the choice of treatment to the clinician. This allows the clinician to recommend treatment based on cost, local availability, patient individual preference, and clinical experience. One way to assess if there is sufficient evidence to recommend a unique optimal treatment for a patient is to construct a confidence interval for the predicted difference in mean response across treatments. In the case of binary treatments, for a fixed patient history $H_t = h_t$, one would construct a confidence interval for the difference $Q_t(h_t, 1; \beta_t^*) - Q_t(h_1, -1; \beta_t^*) = c^\intercal \beta_t^*$ where $c = (\mathbf{0}^\intercal, h_{t,1}^\intercal)^\intercal$. If this confidence interval contains zero then one would conclude that there is insufficient evidence at the nominal level for a unique best treatment.

In this example, the patient features that interact with treatment are categorical. Consequently, we can construct confidence intervals for the predicted difference in mean response across treatments for every possible patient history. These confidence intervals are given in table (3.4). The 90% confidence intervals suggest that there is insufficient evidence at the first stage to recommend a unique best treatment for each patient history. Rather, we would prefer not to make a strong recommendation at stage one, and leave treatment choice solely at the discretion of the clinician. Conversely, in the second stage, the 90% confidence intervals suggest that there is substantial evidence to recommend a unique best treatment when a patient had low adherence—knowledge that is important for evidence-based clinical decision making.

Stage	History	Contrast for $\beta_{t,1}$	Lower (5%)	Upper (95%)	Conclusion
1	Had prior med.	(1 1)	-0.49	0.14	Insufficient evidence
1	No prior med.	(1 0)	-0.05	0.38	Insufficient evidence
2	High adherence and BMOD	(1 1 1)	-0.08	0.69	Insufficient evidence
2	Low adherence and BMOD	(1 0 1)	-1.10	-0.28	Sufficient evidence
2	High adherence and MEDS	(1 1 -1)	-0.18	0.62	Insufficient evidence
2	Low adherence and MEDS	(1 0 -1)	-1.25	-0.29	Sufficient evidence

Table 3.11: Confidence intervals for the predicted difference in mean response across treatments across different patient histories at the 90% level. Confidence intervals that do not contain zero are deemed as having insufficient evidence for recommending a unique best treatment.

3.5 Discussion

The task of constructing valid confidence intervals for the parameters in the Q-function is both scientifically important and statistically challenging. In this paper we offer a first step toward conducting inference in DTRs that is theoretically sound, computationally efficient, and easy to apply. The method presented here provides asymptotically valid intervals regardless of the true configuration of underlying parameters β_t^* or the joint distribution of patient histories H_t for $t = 1, 2, \ldots, T$. Theoretical guarantees were supported by a suite of test examples in which the *ACI* performed favorably to competitors. The ACI is conservative when all of the coefficients of terms involving the second stage treatment are zero. It is our experience that efforts to reduce this conservatism negatively impacts the performance of the resulting confidence interval for other generative models; we conjecture that this conservatism can not be ameliorated without negatively impacting the overall performance of the confidence interval.

There are a number of avenues for future work on this problem. We mention three of the most interesting here. The first is extending to ACI to problems where parameters are shared across stages. This setting occurs when a patients status is modeled as series of renewals (as is often assumed in settings with a very large number of stages) or when smoothness across stages is assumed. Another area of interest is the so-called "large p small n" paradigm where the number of predictors in the Q-function far exceeds the number of observations. This setting arises, for example, when a patients genetic information might be used to tailor treatment. A complication to question of inference in this setting is that it is preceded by the more fundamental question how one should even build Q-functions in this setting. Penalized estimation and Q-learning in one stage decision problems are discussed in (Qian and Murphy 2009)and in multi-stage problems in (Song et al., 2010). A last area of interest is that of reducing the bias in the estimation of the stage 1 treatment effect (recall that if the stage 2 effect is zero for a some patient features then the bias is of order $1/\sqrt{n}$). The most promising work in this area seems to be that of Song et al., (2010) although this work induces additional non-regularity; it would be most interesting to develop confidence intervals that reflect the variability due to the variable selection used in Song et al.

3.6 Appendix I: Proofs for the ACI for two stages and two treatments per stage

The following results will be used repeatedly in the following proofs.

Lemma 3.6.1 (Trivial inequality)**.** *Let a, b be scalars and $[.]_+$ denote the map $z \mapsto \max(0, z)$ then*

$$[a+b]_+ - [b]_+ \leq [a]_+ \leq |a|.$$

Proof. The first inequality follows since $[.]_+$ is monotone non-decreasing, and second inequality follows since $[.]_+$ is bounded above by $|.|$. □

Corollary 3.6.2 (Lipschitz Continuity)**.** *The function* $[.]_+$ *is Lipschitz continuous with Lipschitz constant* 1*.*

Proof. Notice that for any scalars a and b we have

$$[a]_+ - [b]_+ = [(a-b)+b]_+ - [b]_+ \leq [a-b]_+ \leq |a-b|,$$

since we can interchange the role of a and b the result follows. □

Proof of Remark 2.1. We begin by proving part 1 of the remark. By construction $\sqrt{n}\beta^*_{2,1}$ belongs to $\mathcal{S}_n$ for all n and hence $\mathcal{S}_n$ satisfies Condition I. To see that $\mathcal{S}_n$ satisfies condition II, let γ_n be a sequence in $\mathbb{R}^{\dim(\beta^*_{2,1})}$ satisfying $||\mathbb{P}_{\mathcal{H}}\gamma_n||^2_\infty/\lambda_n \to 0$ as $n \to \infty$. Recall that $\mathcal{H}$ is the hyperplane in $\mathbb{R}^{\dim(\beta^*_{2,1})}$ with normal vector $\beta^*_{2,1}$, and $\mathbb{P}_{\mathcal{H}}$ denotes the orthogonal projection onto $\mathcal{H}$. Then, let $v_n \triangleq \sqrt{n}\beta^*_{2,1} + \mathbb{P}_{\mathcal{H}}\gamma_n$ and notice that $||v_n - \sqrt{n}\beta^*_{2,1}||^2_\infty = ||\mathbb{P}_{\mathcal{H}}\gamma_n||^2_\infty$ which is less than λ_n for sufficiently large n. Thus, v_n belongs to $\mathcal{S}_n$ for sufficiently large n and since $\mathbb{P}_{\mathcal{H}}v_n = \mathbb{P}_{\mathcal{H}}\gamma_n$ it follows that $\mathcal{S}_n$ satisfies Condition II.

To prove part 2 of the remark assume $\mathcal{S}'_n$ satisfies Conditions I and II. It is easily seen that $\text{diam}(\mathbb{P}_{\mathcal{H}}\mathcal{S}_n) = O(\sqrt{\lambda_n})$. Suppose towards a contradiction that $\text{diam}(\mathbb{P}_{\mathcal{H}}\mathcal{S}'_n) = O(\sqrt{\tau_n\lambda_n})$ for some sequence of positive scalars τ_n which satisfy $\tau_n \to 0$ as n tends to ∞. Let u_n be a sequence of positive scalars converging to zero as n tends to ∞ satisfying $u_n = o(\tau_n)$. Further, let $\mathcal{B}(\sqrt{u_n\lambda_n}, 0)$ be a ball of radius $\sqrt{u_n\lambda_n}$ in $\mathbb{R}^{\dim(\beta^*_{2,1})}$. Then any sequence of vectors γ_n in $\mathcal{B}(\sqrt{u_n\lambda_n}, 0)$ satisfies $||\mathbb{P}_{\mathcal{H}}\gamma_n||^2_\infty/\lambda_n \to 0$. Consequently, it must be the case $\text{diam}(\mathbb{P}_n\mathcal{S}'_n) = O(\sqrt{u_n\lambda_n})$ which is a contradiction since τ_n/u_n diverges to ∞. □

Lemma 3.6.3 (Consistent Treatment Assignment)**.** *Assume (A1)-(A3). Then*

$$P||B_1||\,||H_{2,1}||1_{(H_{2,1}^\intercal\hat{\beta}_{2,1})(H_{2,1}^\intercal\beta_{2,1}^*)<0} = o_P(1).$$

Proof. We can decompose $P||B_1||\,||H_{2,1}||1_{(H_{2,1}^\intercal\hat{\beta}_{2,1})(H_{2,1}^\intercal\beta_{2,1}^*)<0}$ into

$$P||B_1||\,||H_{2,1}||1_{H_{2,1}^\intercal\hat{\beta}_{2,1}<0}1_{H_{2,1}^\intercal\beta_{2,1}^*>0} + P||B_1||\,||H_{2,1}||1_{H_{2,1}^\intercal\hat{\beta}_{2,1}>0}1_{H_{2,1}^\intercal\beta_{2,1}^*<0}.$$

It is sufficient to show that the first term in the above decomposition is $o_P(1)$ since the same argument will hold if we replace $H_{2,1}$ with $-H_{2,1}$. For any $\epsilon > 0$ we can write this term as

$$P||B_1||\,||H_{2,1}||1_{H_{2,1}^\intercal\hat{\beta}_{2,1}<0}1_{\frac{H_{2,1}^\intercal\beta_{2,1}^*}{||H_{2,1}||}>\epsilon} + P||B_1||\,||H_{2,1}||1_{H_{2,1}^\intercal\hat{\beta}_{2,1}<0}1_{0<\frac{H_{2,1}^\intercal\beta_{2,1}^*}{||H_{2,1}||}\le\epsilon}. \tag{3.20}$$

Note that we need not worry about the case that $H_{2,1} = 0$ since this is clearly not included in the event $(H_{2,1}^\intercal\hat{\beta}_{2,1})(H_{2,1}^\intercal\beta_{2,1}^*) < 0$. The second term in (3.20) is bounded above by

$$P||B_1||\,||H_{2,1}||1_{0<\frac{H_{2,1}^\intercal\beta_{2,1}^*}{||H_{2,1}||}\le\epsilon}, \tag{3.21}$$

which, by (A1), can be made arbitrarily small for sufficiently small ϵ. Let $\eta > 0$ be arbitrary and choose $\epsilon > 0$ so that (3.21) is smaller than $\eta/2$. Hereafter, we regard ϵ as fixed. Turning attention to the first term in (3.20) we note that the event

$$\left\{H_{2,1}^\intercal\hat{\beta}_{2,1} < 0\,,\ \frac{H_{2,1}^\intercal\beta_{2,1}^*}{||H_{2,1}||} > \epsilon\right\}$$

is equal to

$$\left\{\frac{H_{2,1}^\intercal\mathbb{V}_n}{||H_{2,1}||} < -\frac{H_{2,1}^\intercal\beta_{2,1}^*\sqrt{n}}{||H_{2,1}||}\,,\ \frac{H_{2,1}^\intercal\beta_{2,1}^*}{||H_{2,1}||} > \epsilon\right\}$$

which is contained in

$$\left\{ \frac{H_{2,1}^{\intercal}\mathbb{V}_n}{||H_{2,1}||} \leq -\epsilon\sqrt{n} \right\} \subseteq \left\{ -\frac{|H_{2,1}^{\intercal}\mathbb{V}_n|}{||H_{2,1}||} < -\epsilon\sqrt{n} \right\} \subseteq \left\{ ||\mathbb{V}_n|| > \epsilon\sqrt{n} \right\}.$$

where the last containment follows from the Cauchy-Schwartz inequality. The preceding series of containments show that

$$P||B_1||\,||H_{2,1}||1_{H_{2,1}^{\intercal}\hat{\beta}_{2,1}<0}1_{\frac{H_{2,1}^{\intercal}\beta_{2,1}^{*}}{||H_{2,1}||}>\epsilon} \leq 1_{||\mathbb{V}_n||>\epsilon\sqrt{n}}P||B_1||\,||H_{2,1}|| = o_p(1).$$

Let $\delta > 0$ be arbitrary and keep $\epsilon > 0$ fixed, from the above we can choose n sufficiently large so that

$$P\left(P||B_1||\,||H_{2,1}||1_{H_{2,1}^{\intercal}\hat{\beta}_{2,1}<0}1_{\frac{H_{2,1}^{\intercal}\beta_{2,1}^{*}}{||H_{2,1}||}>\epsilon} > \eta/2 \right) \leq \delta,$$

where the outer probability statement is over training sets of size n (e.g. the probability statement applies to $\hat{\beta}_{2,1}$). Putting all of the above together, it follows that

$$\begin{aligned} &P\left(P||B_1||\,||H_{2,1}||1_{H_{2,1}^{\intercal}\hat{\beta}_{2,1}<0}1_{H_{2,1}^{\intercal}\beta_{2,1}^{*}>0} > \eta \right) \\ &\leq 1_{P||B_1||\,||H_{2,1}||1_{0<\frac{H_{2,1}^{\intercal}\beta_{2,1}^{*}}{||H_{2,1}||}\leq\epsilon}>\eta/2} + P\left(P||B_1||\,||H_{2,1}||1_{H_{2,1}^{\intercal}\hat{\beta}_{2,1}<0}1_{\frac{H_{2,1}^{\intercal}\beta_{2,1}^{*}}{||H_{2,1}||}>\epsilon} > \eta/2 \right) \\ &= P\left(P||B_1||\,||H_{2,1}||1_{H_{2,1}^{\intercal}\hat{\beta}_{2,1}<0}1_{\frac{H_{2,1}^{\intercal}\beta_{2,1}^{*}}{||H_{2,1}||}>\epsilon} > \eta/2 \right) \leq \delta, \end{aligned}$$

where the equality holds since the indicator is identically zero by our choice of ϵ. Since δ and η were arbitrary the result is proved. □

Proof of Theorem 2.1, Part 1. Here, we derive the limiting distribution of $c^{\intercal}\sqrt{n}(\hat{\beta}_1 -$

$\beta_1^*)$ under fixed alternatives. Recall that the following decomposition holds

$$c^\intercal\sqrt{n}(\hat{\beta}_1 - \beta_1^*) = c^\intercal\mathbb{W}_n + c^\intercal\Sigma_{1,n}^{-1}\mathbb{P}_n B_1^\intercal\mathbb{U}_n, \tag{3.22}$$

where

$$\begin{aligned}\mathbb{W}_n &= \Sigma_{1,n}^{-1}\sqrt{n}\mathbb{P}_n B_1^\intercal\Big[\left(Y_1 + H_{2,0}^\intercal\beta_{2,0}^* + \left[H_{2,1}^\intercal\beta_{2,1}^*\right]_+ - B_1\beta_1^*\right) + H_{2,0}^\intercal\left(\hat{\beta}_{2,0} - \beta_{2,0}^*\right)\Big],\\ \mathbb{U}_n &= \sqrt{n}\left(\left[H_{2,1}^\intercal\hat{\beta}_{2,1}\right]_+ - \left[H_{2,1}^\intercal\beta_{2,1}^*\right]_+\right).\end{aligned}$$

The term $\mathbb{W}_n$ is asymptotically Gaussian with mean zero since, using the definition of β_1^*, we can express it as

$$\Sigma_{1,n}^{-1}\sqrt{n}(\mathbb{P}_n - P)B_1^\intercal\left(Y_1 + H_{2,0}^\intercal\beta_{2,0}^* + \left[H_{2,1}^\intercal\beta_{2,1}^*\right]_+ - B_1\beta_1^*\right) + \Sigma_{1,n}^{-1}\mathbb{P}_n H_{2,0}^\intercal\sqrt{n}\left(\hat{\beta}_{2,0} - \beta_{2,0}^*\right),$$

which is asymptotically normal by the multivariate central limit theorem and Slutsky's Theorem (we have used (A1) and (A2) here). It follows that the second term in the decomposition of $c^\intercal\sqrt{n}(\hat{\beta}_1 - \beta_1^*)$ is equal to $c^\intercal\Sigma_{1,n}^{-1}PB_1^\intercal\mathbb{U}_n + o_P(1)$. To see this, recall that in the main body of the paper we defined $\mathbb{V}_n \triangleq \sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1}^*)$. From standard results for least squares estimators, it is seen that $\mathbb{V}_n$ is asymptotically Gaussian with mean zero. Notice that

$$\mathbb{U}_n = \left[H_{2,1}^\intercal\left(\mathbb{V}_n + \sqrt{n}\beta_{2,1}^*\right)\right]_+ - \left[\sqrt{n}H_{2,1}^\intercal\beta_{2,1}^*\right]_+. \tag{3.23}$$

Thus, using the inequality (3.6.1) and the Cauchy-Schwartz inequality, we see that

$$\left|\left|\mathbb{P}_n B_1^\intercal\mathbb{U}_n\right|\right| \leq \mathbb{P}_n||B_1||\ \left[H_{2,1}^\intercal\mathbb{V}_n\right]_+ \leq \mathbb{P}_n||B_1||\left|H_{2,1}^\intercal\mathbb{V}_n\right| \leq \mathbb{P}_n||B_1||\ ||H_{2,1}||\ ||\mathbb{V}_n|| = O_P(1),$$

where the last equality follows from the asymptotic normality of $\mathbb{V}_n$ and the LLN. Slutsky's theorem and the convergence in probability of $\Sigma_{1,n}$ to the positive definite matrix $\Sigma_{1,\infty} \triangleq PB_1^\intercal B_1$, allows us to write

$$c^\intercal \Sigma_{1,n}^{-1} \mathbb{P}_n \mathbb{U}_n = c^\intercal \Sigma_{1,\infty}^{-1} \mathbb{P}_n \mathbb{U}_n + o_P(1).$$

Now, write the second term in (3.22) as

$$c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal \mathbb{U}_n + c^\intercal \Sigma_{1,\infty} (\mathbb{P}_n - P) B_1^\intercal \mathbb{U}_n + o_P(1).$$

Let ϵ and η be arbitrary positive constants it suffices to show that

$$P\left(\left|c^\intercal \Sigma_{1,\infty}^{-1} (\mathbb{P}_n - P) B_1^\intercal \mathbb{U}_n\right| > \epsilon\right) \leq \eta \tag{3.24}$$

for all n sufficiently large. For arbitrary $M > 0$ we have

$$\begin{aligned} c^\intercal \Sigma_{1,\infty}^{-1} (\mathbb{P}_n - P) B_1^\intercal \mathbb{U}_n = {} & c^\intercal \Sigma_{1,\infty}^{-1} (\mathbb{P}_n - P) B_1^\intercal \left[H_{2,1}^\intercal \left(\mathbb{V}_n + \sqrt{n}\beta_{2,1}^*\right)\right]_+ - \left[\sqrt{n} H_{2,1}^\intercal \beta_{2,1}^*\right]_+ 1_{||\mathbb{V}_n|| \leq M} \\ & + c^\intercal \Sigma_{1,\infty}^{-1} (\mathbb{P}_n - P) B_1^\intercal \left[H_{2,1}^\intercal \left(\mathbb{V}_n + \sqrt{n}\beta_{2,1}^*\right)\right]_+ - \left[\sqrt{n} H_{2,1}^\intercal \beta_{2,1}^*\right]_+ 1_{||\mathbb{V}_n|| > M}. \end{aligned}$$

Thus, the requisite probability in (3.24) is bounded above by

$$\begin{aligned} P\left(\sup_{\gamma,\delta \in \mathbb{R}^{\dim(\beta_{2,1}^*)}} \left| c^\intercal \Sigma_{1,\infty}^{-1} \left(\mathbb{P}_n - P\right) B_1^\intercal \left(\left[H_{2,1}^\intercal (\gamma + \delta)\right]_+ - \left[H_{2,1}^\intercal \delta\right]_+\right) 1_{||\gamma|| \leq M} \right| \geq \epsilon/2\right) \\ + P\left(||\mathbb{V}_n|| > M\right). \end{aligned}$$

Since $\mathbb{V}_n$ converges in distribution, we can select M so that the second probability is less than $\eta/2$ for all n. Let c be an arbitrary fixed vector in $\mathbb{R}^{\dim(B_1)}$. Notice that the class of functions

$$\mathcal{F} \triangleq \left\{ c^{\intercal} B_1^{\intercal} \left(\left[H_{2,1}^{\intercal}(\gamma + \delta) \right]_+ - \left[H_{2,1}^{\intercal} \delta \right]_+ \right) \,:\, \delta \in \mathbb{R}^{\dim(\beta_{2,1}^*)}, \gamma \in \mathcal{B}(M, 0) \right\}$$

is a P-measurable, Bounded Uniform Entropy Integral (BUEI) class with square-integrable envelope function and is therefore Glivenko-Cantelli (Kosorok 2008). Consequently, the probability in (3.24) can be made sufficiently small.

That $\mathcal{F}$ is $BUEI$ follows by noting first that the simpler class

$$\mathcal{F}_1 \triangleq \left\{ H_{2,1}^{\intercal} \eta \,,\, \eta \in \mathbb{R}^{\dim(\beta_{2,1}^*)} \right\}$$

is a VC class with dimension less than or equal to $\dim(\beta_{2,1}^*) + 2$ (see van der Vaart and Wellner 1996, Lemma 2.6.15). The function $[.]_+$ is monotone and hence $[\mathcal{F}_1]_+ \triangleq \{[f]_+ : f \in \mathcal{F}_1\}$ is also VC (see van der Vaart and Wellner 1996, Lemma 2.6.18). For any two classes of functions, say $\mathcal{G}$ and $\mathcal{H}$ we write $\mathcal{G} - \mathcal{H}$ to mean the new class of functions $\{g - h \,:\, g \in \mathcal{G} \,,\, h \in \mathcal{H}\}$. Permanence properties of VC classes (see for example, van der Vaart and Wellner 1996; Dudley 1999; Kosorok 2008) imply that if we let $\mathcal{F}_1'$ be an independent copy of $\mathcal{F}_1$, then $c^{\intercal} B_1^{\intercal}(\mathcal{F}_1 - \mathcal{F}_1')$ is also a VC class. Finally, since $\mathcal{F}$ is a subset of $c^{\intercal} B_1^{\intercal}(\mathcal{F}_1 - \mathcal{F}_1')$, it must also be VC. Any VC class automatically satisfies the conditions to be a Uniform Entropy Integral class (see Kosorok 2008, Theorem 9.3). Lastly, using (3.6.1) we see that $\mathcal{F}$ is bounded by envelope $||B_1||\, ||H_{2,1}||\, M$ which is square-integrable by (A1). Thus, the class $\mathcal{F}$ is BUEI. The class $\mathcal{F}$ can be seen to be P-measurable by noting that it satisfies the stronger condition of being pointwise measurable since any function in $\mathcal{F}$ can be arbitrarily closely approximated by a function in

$$\mathcal{G} \triangleq \left\{ c^\intercal B_1^\intercal \left(\left[H_{2,1}^\intercal(\gamma+\delta) \right]_+ - \left[H_{2,1}^\intercal \delta \right]_+ \right) \,:\, \delta \in \mathbb{Q}^{\dim(\beta_{2,1}^*)}, \gamma \in \mathcal{B}(M,0) \cap \mathbb{Q}^{\dim(\beta_{2,1}^*)} \right\},$$

where $\mathbb{Q}$ denotes the rational numbers (see Kosorok 2008 for a discussion of pointwise measurability).

As our final step in dealing with the second term of (3.22), we will make use of the following decomposition

$$c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal \mathbb{U}_n = c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal \mathbb{U}_n 1_{H_{2,1}^\intercal \beta_{2,1}^* \neq 0} + c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal \left[H_{2,1}^\intercal \mathbb{V}_n \right]_+ 1_{H_{2,1}^\intercal \beta_{2,1}^* = 0}. \quad (3.25)$$

The first term in the above decomposition can be seen to equal

$$\left(c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal H_{2,1}^\intercal 1_{H_{2,1}^\intercal \beta_{2,1}^* > 0} \right) \mathbb{V}_n + o_P(1). \quad (3.26)$$

To see this, first write

$$c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal \mathbb{U}_n 1_{H_{2,1}^\intercal \beta_{2,1}^* \neq 0} = c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal \mathbb{U}_n 1_{(H_{2,1}^\intercal \hat{\beta}_{2,1})(H_{2,1}^\intercal \beta_{2,1}^*) > 0} + c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal \mathbb{U}_n 1_{(H_{2,1}^\intercal \hat{\beta}_{2,1})(H_{2,1}^\intercal \beta_{2,1}^*) < 0},$$

and note that the second term is $o_p(1)$ by appeal to Lemma's (3.6.1) and (3.6.3). Furthermore,

$$c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal \mathbb{U}_n 1_{(H_{2,1}^\intercal \hat{\beta}_{2,1})(H_{2,1}^\intercal \beta_{2,1}^*) > 0} = c^\intercal \Sigma_{1,\infty}^{-1} P B_1^\intercal \mathbb{V}_n 1_{H_{2,1}^\intercal \hat{\beta}_{2,1} > 0} 1_{H_{2,1}^\intercal \beta_{2,1}^* > 0}.$$

Writing $1_{H_{2,1}^\intercal \hat{\beta}_{2,1} > 0} = 1 - 1_{H_{2,1}^\intercal \hat{\beta}_{2,1} < 0}$ and applying Lemma (3.6.3) again gives the result. The final step of the proof is to use the joint asymptotic normality of $\mathbb{W}_n$ and $\mathbb{V}_n$ (this follows from (A1) and (A2)) coupled with the continuous mapping theorem. □

Before providing the proof of the second part of Theorem 2.1, it will be convenient to introduce the following lemma.

Lemma 3.6.4. *Assume (A1), (A2), and (A4). It follows that*

$$\left|\left|\sqrt{n}(\mathbb{P}_n - P_{n,\gamma})B_1^\intercal \left(\left[H_{2,1}^\intercal(\beta_{2,1}^* + \eta)\right]_+ - \left[H_{2,1}^\intercal\beta_{2,1}^*\right]_+\right)\right|\right| = o_{P_{n,\gamma}}(1)$$

uniformly over η in compact sets.

Proof. Fix a vector $c \in \mathbb{R}^{\dim(B_1)}$ with $||c|| = 1$, then it suffices to show that

$$\sqrt{n}(\mathbb{P}_n - P_{n,\gamma})c^\intercal B_1^\intercal \left(\left[H_{2,1}^\intercal(\beta_{2,1}^* + \eta)\right]_+ - \left[H_{2,1}^\intercal\beta_{2,1}^*\right]_+\right) = o_{P_{n,\gamma}}(1)$$

uniformly over η in compact sets. Notice that the class

$$\mathcal{F} \triangleq \left\{f(H_{2,1};\eta) = c^\intercal B_1^\intercal \left(\left[H_{2,1}^\intercal(\beta_{2,1}^* + \eta\right]_+ - \left[H_{2,1}^\intercal\beta_{2,1}^*\right]_+\right) \;\; \eta \in \mathcal{B}(M,0)\right\}$$

is a P-measurable BUEI class which has, by appeal to Lemma (3.6.1), square integrable envelope $||B_1||\,||H_{2,1}||$. That $\mathcal{F}$ is BUEI follows from the fact that $H_{2,1}^\intercal(\beta_{2,1}^*+\eta)$ is BUEI over $\eta \in \mathcal{B}(M,0)$ which implies the monotone transformation $z \mapsto [z]_+$ of $H_{2,1}^\intercal(\beta_{2,1}^* + \eta)$ is also BUEI (see Kosorok 2008). Finally, translating and scaling by functions that do not depend on η is BUEI preserving. It follows from Theorem 11.12 of Kosorok (2008) that

$$\sqrt{n}(\mathbb{P}_n - P_{n,\gamma}) \rightsquigarrow \mathbb{G} \quad \in l^\infty(\mathcal{F}),$$

where $\mathbb{G}$ is a Brownian bridge with covariance function given by

$$Cov(\mathbb{G}(\delta), \mathbb{G}(\eta)) =$$

$$P\left(c^{\intercal}B_1^{\intercal}\left(\left[H_{2,1}^{\intercal}(\beta_{2,1}^{*}+\delta)\right]_+ - \left[H_{2,1}^{\intercal}\beta_{2,1}^{*}\right]_+\right) - Pc^{\intercal}B_1^{\intercal}\left(\left[H_{2,1}^{\intercal}(\beta_{2,1}^{*}+\delta)\right]_+ - \left[H_{2,1}^{\intercal}\beta_{2,1}^{*}\right]_+\right)\right)$$
$$\times\left(c^{\intercal}B_1^{\intercal}\left(\left[H_{2,1}^{\intercal}(\beta_{2,1}^{*}+\eta)\right]_+ - \left[H_{2,1}^{\intercal}\beta_{2,1}^{*}\right]_+\right) - Pc^{\intercal}B_1^{\intercal}\left(\left[H_{2,1}^{\intercal}(\beta_{2,1}^{*}+\eta)\right]_+ - \left[H_{2,1}^{\intercal}\beta_{2,1}^{*}\right]_+\right)\right).$$

The limiting process $\mathbb{G}$ is stochastically equicontinuous with respect to the covariance pseudo metric $\rho(\delta,\eta)^2 \triangleq P\left(\mathbb{G}(\delta)-\mathbb{G}(\eta)\right)^2$ (see van der Vaart and Wellner 1996 or Kosorok 2008). The result follows if we can show that $||\delta||^2 \to 0$ implies that $\rho(\delta, 0) \to 0$. But this comes out of the form of the covariance function of $\mathbb{G}$ and noting that

$$Pc^{\intercal}B_1^{\intercal}\left(\left[H_{2,1}^{\intercal}(\beta_{2,1}^{*}-\delta)\right]_+ - \left[H_{2,1}^{\intercal}\beta_{2,1}^{*}\right]_+\right) \to 0$$

as δ tends to zero. □

Proof of Theorem 2.1, Part 2. Assume (A1)-(A4). Our goal is to derive the limiting distribution of $c^{\intercal}\sqrt{n}(\hat{\beta}_1 - \delta_1^{(n)})$ under $P_{n,\gamma}$, where

$$\delta_1^{(n)} \triangleq \arg\min_{\delta} P_{n,\gamma}(\tilde{Y}_{1,n}^{*} - Q_1(H_1, A_1; \delta))^2,$$

and

$$\tilde{Y}_{1,n}^{*} \triangleq Y_1 + H_{2,0}^{\intercal}\delta_{2,0}^{(n)} + \left[H_{2,1}^{\intercal}\delta_{2,1}^{(n)}\right]_+.$$

From the normal equations and the definition of the predicted future reward following the estimated optimal policy $\tilde{Y}_{1,n}^{*}$, it follows that

$$\begin{aligned}\sqrt{n}(\hat{\beta}_1 - \delta_1^{(n)}) &= \Sigma_{1,n}^{-1}\sqrt{n}\mathbb{P}_n B_1^{\intercal}\left(\tilde{Y}_{1,n}^{*} - B_1\delta_1^{(n)}\right)\\ &= \Sigma_{1,n}^{-1}\sqrt{n}\mathbb{P}_n B_1^{\intercal}\left(Y_1 + H_{2,0}^{\intercal}\hat{\beta}_{2,0} + \left[H_{2,1}^{\intercal}\hat{\beta}_{2,1}\right]_+ - B_1\delta_1^{(n)}\right).\end{aligned}$$

Adding and subtracting terms to the right hand side of the above display, we can further decompose $\sqrt{n}(\hat{\beta}_1 - \delta_1^{(n)})$ as

$$\begin{aligned}\Sigma_{1,n}^{-1}\sqrt{n}(\mathbb{P}_n - P_{n,\gamma})\left(Y_1 + H_{2,0}^{\intercal}\delta_{2,0}^{(n)} + \left[H_{2,1}^{\intercal}\delta_{2,1}^{(n)}\right]_+ - B_1\delta_1^{(n)}\right) + \Sigma_{1,n}^{-1}\mathbb{P}_n B_1^{\intercal}H_{2,0}^{\intercal}\sqrt{n}\left(\hat{\beta}_{2,0} - \delta_{2,0}^{(n)}\right) \\ + \Sigma_{1,n}^{-1}\sqrt{n}\mathbb{P}_n B_1^{\intercal}\left(\left[H_{2,1}^{\intercal}\hat{\beta}_{2,1}\right]_+ - \left[H_{2,1}^{\intercal}\delta_{2,1}^{(n)}\right]_+\right)1_{H_{2,1}^{\intercal}\beta_{2,1}^{*}\neq 0} \\ + \Sigma_{1,n}^{-1}\mathbb{P}_n B_1^{\intercal}\left(\left[H_{2,1}^{\intercal}(\sqrt{n}(\hat{\beta}_{2,1} - \delta_{2,1}^{(n)}) + \gamma)\right]_+ - \left[H_{2,1}^{\intercal}\gamma\right]_+\right)1_{H_{2,1}^{\intercal}\beta_{2,1}^{*}=0}.\end{aligned} \tag{3.27}$$

We have assumed that $\delta_{2,1}^{(n)} = \beta_{2,1}^{*} + \gamma/\sqrt{n}$, owing to (A4) it must also be the case that $\delta_1^{(n)} = \beta_1^{*} + O(1/\sqrt{n})$ and $\delta_{2,0}^{(n)} = \beta_{2,0}^{*} + O(1/\sqrt{n})$. To see this, note from the definition of $\delta_1^{(n)}$ it follows that $\delta_1^{(n)}$ must satisfy

$$P_{n,\gamma}B_1^{\intercal}\left(\tilde{Y}_{1,n}^{*} - B_1\delta_1^{(n)}\right) = 0$$

so that for sufficiently large n, $\delta_1^{(n)} = \left(P_{n,\gamma}B_1^{\intercal}B_1\right)^{-1}P_{n,\gamma}B_1^{\intercal}\tilde{Y}_{1,n}^{*}$. Furthermore, we have

$$\begin{aligned}\left(P_{n,\gamma}B_1^{\intercal}B_1\right)^{-1}P_{n,\gamma}B_1^{\intercal}\tilde{Y}_{1,n}^{*} &= \left(PB_1^{\intercal}B_1 + (P_{n,\gamma} - P)B_1^{\intercal}B_1\right)^{-1}\left(PB_1^{\intercal}\tilde{Y}_{1,n}^{*} + (P_{n,\gamma} - P)B_1^{\intercal}\tilde{Y}_{1,n}^{*}\right) \\ &= \left(PB_1^{\intercal}B_1 + O(1/\sqrt{n})\right)^{-1}\left(PB_1^{\intercal}\tilde{Y}_{1,n}^{*} + O(1/\sqrt{n})\right),\end{aligned} \tag{3.28}$$

where the last equality follows from Theorem 11.12 in Kosorok (2008). Since the eigenvalues of $PB_1^{\intercal}B_1$ are bounded away from zero, we can re-write the last term in (3.28) as

$$(PB_1^{\intercal}B_1)^{-1}\left(PB_1^{\intercal}\tilde{Y}_{1,n}^{*}\right) + O(1/\sqrt{n}).$$

This means that $\delta_1^{(n)} = \beta_1^{*} + O(1/\sqrt{n})$, since $\tilde{Y}_{1,n}^{*} = Y_1 + H_{2,0}^{\intercal}\beta_{2,0}^{*} + \left[H_{2,1}^{\intercal}\beta_{2,1}^{*}\right]_+ + O(1/\sqrt{n})$, and $\beta_1^{*} = (PB_1^{\intercal}B_1)^{-1}PB_1^{\intercal}\left(Y_1 + H_{2,0}^{\intercal}\beta_{2,0}^{*} + \left[H_{2,1}^{\intercal}\beta_{2,1}^{*}\right]_+\right)$. Similarly, we have $\delta_2^{(n)} = (P_{n,\gamma}B_2^{\intercal}B_2)^{-1}(P_{n,\gamma}B_2^{\intercal}Y_2)$ which, by an identical argument is equal to $(PB_2^{\intercal}B_2)^{-1}(PB_2^{\intercal}Y_2) + O(1/\sqrt{n}) = \beta_2^{*} + O(1/\sqrt{n})$.

The preceding discussion and Lemma (3.6.4) show that under (A1)-(A4) the first term in (3.27) is equal to

$$\Sigma_{1,n}^{-1}\sqrt{n}(\mathbb{P}_n - P_{n,\gamma})B_1^{\intercal}\left(Y_1 + H_{2,0}^{\intercal}\beta_{2,0}^{*} + \left[H_{2,1}^{\intercal}\beta_{2,1}^{*}\right]_{+} - B_1\beta_1^{*}\right) + o_{P_{n,\gamma}}(1). \tag{3.29}$$

Furthermore,

$$\begin{aligned}\sqrt{n}(\hat{\beta}_2 - \delta_2^{(n)}) &= \Sigma_{2,n}^{-1}\sqrt{n}\mathbb{P}_n\left(B_2^{\intercal}Y_2 - \Sigma_{2,n}(P_{n,\gamma}B_2^{\intercal}B_2)^{-1}P_{n,\gamma}B_2^{\intercal}Y_2\right)\\ &= \Sigma_{2,n}^{-1}\sqrt{n}\mathbb{P}_n\left(B_2^{\intercal}Y_2 - (\mathbf{I} + o_{P_{n,\gamma}}(1))P_{n,\gamma}B_2^{\intercal}Y_2\right)\\ &= \Sigma_{2,n}^{-1}\sqrt{n}(\mathbb{P}_n - P_{n,\gamma})B_2^{\intercal}Y_2 + o_{P_{n,\gamma}}(1).\end{aligned}$$

The conclusion of the preceding set of equalities is that upon applying Theorem 11.12 of Kosorok (2008) it follows that $\sqrt{n}(\hat{\beta}_2 - \delta_2^{(n)})$ has the same limiting distribution under $P_{n,\gamma}$ as $\sqrt{n}(\hat{\beta}_2 - \beta_2^{*})$ does under P. Combining this result with (3.29) it follows that the first two terms in (3.27) converge jointly to $\mathbb{W}_\infty$ under $P_{n,\gamma}$. Thus, the local parameter does not appear in this component of the limiting distribution of $\sqrt{n}(\hat{\beta}_1 - \delta_1^{(n)})$ associated with the first two terms of (3.27). However, as seen in the statement of Theorem 2.1., the local parameter will appear in the limiting distribution of the last term of (3.27).

We now derive the limiting distribution of the last two terms of (3.27). The same techniques employed in the proof of part 1 of Theorem 2.1. can be used to show that the last two terms are equal to

$$\begin{aligned}&\Sigma_{1,\infty}^{-1}\mathbb{P}_n B_1^{\intercal}H_{2,1}^{\intercal}1_{H_{2,1}^{\intercal}\beta_{2,1}^{*}>0}\sqrt{n}(\hat{\beta}_{2,1} - \delta_{2,1}^{(n)})\\ &\quad + \Sigma_{1,\infty}^{-1}\mathbb{P}_n B_1^{\intercal}\left(\left[H_{2,1}^{\intercal}\left(\sqrt{n}(\hat{\beta}_{2,1} - \delta_{2,1}^{(n)}) + \gamma\right)\right]_{+} - \left[H_{2,1}^{\intercal}\gamma\right]_{+}\right)1_{H_{2,1}^{\intercal}\beta_{2,1}^{*}=0} + o_{P_{n,\gamma}}(1).\end{aligned} \tag{3.30}$$

Lastly, making the substitution $\mathbb{P}_n = P + (P_{n,\gamma} - P) + (\mathbb{P}_n - P_{n,\gamma})$ in (3.30), we see that the last two terms of (3.27) are equal to

$$\Sigma_{1,\infty}^{-1} P B_1^\intercal H_{2,1}^\intercal 1_{H_{2,1}^\intercal \beta_{2,1}^* > 0} \sqrt{n}(\hat{\beta}_{2,1} - \delta_{2,1}^{(n)}) \\ + \Sigma_{1,\infty}^{-1} P B_1^\intercal \left(\left[H_{2,1}^\intercal \left(\sqrt{n}(\hat{\beta}_{2,1} - \delta_{2,1}^{(n)}) + \gamma \right) \right]_+ - \left[H_{2,1}^\intercal \gamma \right]_+ \right) 1_{H_{2,1}^\intercal \beta_{2,1}^* = 0} + o_{P_{n,\gamma}}(1).$$

Application of the continuous mapping theorem and Slutsky's theorem show that the limiting distribution under $P_{n,\gamma}$ of the above quantity is equal to

$$\Sigma_{1,\infty}^{-1} P B_1^\intercal H_{2,1}^\intercal \mathbb{V}_\infty 1_{H_{2,1}^\intercal \beta_{2,1}^* > 0} + \Sigma_{1,\infty}^{-1} P B_1^\intercal \left(\left[H_{2,1}^\intercal (\mathbb{V}_\infty + \gamma) \right]_+ - \left[H_{2,1}^\intercal \gamma \right]_+ \right) 1_{H_{2,1}^\intercal \beta_{2,1}^* = 0}.$$

Note that this convergence is uniform over γ. The theorem is concluded by establishing joint convergence via the Cramer-Wold device and then applying the continuous mapping theorem. □

To prove the third part of Theorem 2.1, the following results will be useful.

Lemma 3.6.5. *Assume (A1)-(A4). For each fixed* $h_{2,1}$*, it follows that* $1_{T_n(h_{2,1}) \le \lambda_n} \to 1_{h_{2,1}^\intercal \beta_{2,1}^* = 0}$ *in probability.*

Proof. Recall that $T_n(h_{2,1}) \triangleq \frac{n(h_{2,1}^\intercal \hat{\beta}_{2,1})^2}{h_{2,1}^\intercal \Sigma_{2,n}^{(2,2)} h_{2,1}}$ so that we can write

$$\frac{T_n(h_{2,1})}{\lambda_n} \triangleq \frac{n(h_{2,1}^\intercal \hat{\beta}_{2,1})^2}{\lambda_n h_{2,1}^\intercal \Sigma_{2,n}^{(2,2)} h_{2,1}} = \frac{(h_{2,1}^\intercal(\mathbb{V}_n/\sqrt{\lambda_n} + \sqrt{n/\lambda_n}\beta_{2,1}^*))^2}{h_{2,1}^\intercal \Sigma_{2,n}^{(2,2)} h_{2,1}}.$$

Noting that $\mathbb{V}_n/\sqrt{\lambda_n} = o_p(1)$, it follows that the last term on the right hand side of the above display tends to $+\infty$ in probality when $h_{2,1}^\intercal \beta_{2,1}^* \neq 0$ but tends to zero when $h_{2,1}^\intercal \beta_{2,1}^* = 0$. This proves the result. □

Corollary 3.6.6. *Assume (A1)-(A4). It follows that both*

1. $P||B_1||\,||H_{2,1}||1_{T_n(H_{2,1})\leq\lambda_n}1_{H_{2,1}^{\intercal}\beta_{2,1}^{*}\neq 0}$

2. $P||B_1||\,||H_{2,1}||1_{T_n(H_{2,1})>\lambda_n}1_{H_{2,1}^{\intercal}\beta_{2,1}^{*}=0}$

converge to zero in probability.

Proof. Let $\gamma_{2,1}^{n}$ be a sequence of constants converging to zero in $\mathbb{R}^{\dim(\beta_{2,1}^{*})}$ and b_n any positive sequence diverging to infinity. Then, for any fixed $c > 0$ the set

$$\left\{h_{2,1}\ :\ h_{2,1}^{\intercal}\beta_{2,1}^{*}\neq 0,\quad \left(\frac{h_{2,1}^{\intercal}\gamma_{2,1}^{n}+b_nh_{2,1}^{\intercal}\beta_{2,1}^{*}}{||h_{2,1}||}\right)^2\leq c\right\}$$

has P measure which is less than the P measure of the set

$$\left(\frac{-\sqrt{c}-||\gamma_n||_\infty}{b_n},0\right)\bigcup\left(0,\frac{\sqrt{c}+||\gamma_n||_\infty}{b_n}\right),$$

where $||.||_\infty$ is the usual supremum norm. The P measure of the set in the above display converges to zero as n tends to infinity. A similar argument shows that the set

$$\left\{h_{2,1}\ :\ h_{2,1}^{\intercal}\beta_{2,1}^{*}=0,\quad \left(\frac{h_{2,1}^{\intercal}\gamma_{2,1}^{n}+b_nh_{2,1}^{\intercal}\beta_{2,1}^{*}}{||h_{2,1}||}\right)^2> c\right\}$$

has P measure converging to zero as n tends to infinity.

Let a_n be an arbitrary monotone increasing sequence of integers, then, there exists a subsequence of a_n, say a_{n_i} so that $\mathbb{V}_{a_{n_i}}/\sqrt{\lambda_{a_{n_i}}}$ converges almost surely to zero and $\Sigma_{2,a_{n_i}}^{(2,2)}$ converges almost surely to $\Sigma_{2,\infty}^{(2,2)}$. Consider $\mathbb{V}_{a_{n_i}}/\sqrt{\lambda_{a_{n_i}}}$ as our $\gamma_{2,1}^{n}$ and $\sqrt{n/\lambda_n}$ as our b_n. The preceding discussion, the proof of Lemma (3.6.5), and the dominated convergence theorem show that

$$P||B_1||\,||H_{2,1}||1_{T_{a_{n_i}}(H_{2,1})\leq\lambda}1_{H_{2,1}^{\intercal}\beta_{2,1}^{*}\neq 0}\to 0$$

almost surely. If we define

$$A_n = P||B_1||\,||H_{2,1}||1_{T_n(H_{2,1})\le\lambda_n}1_{H_{2,1}^\intercal\beta_{2,1}^*\neq 0},$$

then we have shown that for any sequence a_n tending to ∞, there exists a further subsequence, say a_{n_i}, for which $A_{a_{n_i}}$ converges almost surely to zero. Hence, A_n converges in probability to zero. The first part of the corollary is proved. The second part of the corollary follows from an identical argument and is thus omitted. □

Define $PD(\epsilon)$ to be the space of $\mathbb{R}^{\dim(\beta_{2,1}^*)} \times \mathbb{R}^{\dim(\beta_{2,1}^*)}$ positive definite matrices with eigenvalues bounded below by ϵ equipped with the spectral norm.

Lemma 3.6.7. *Assume (A1). Then, the class of functions*

$$\mathcal{F} \triangleq \left\{ ||B_1||\,||H_{2,1}||1_{\frac{(H_{2,1}^\intercal\gamma)^2}{H_{2,1}^\intercal\Omega H_{2,1}}\le\sigma}1_{H_{2,1}^\intercal\beta_{2,1}^*=0} \,:\, \Omega\in PD(\epsilon)\,,\,\sigma\in\mathbb{R}_+\,,\,\gamma\in\mathbb{R}^{\dim(\beta_{2,1}^*)}\right\}$$

is a P-measurable BUEI class with square-integrable envelope $||B_1||\,||H_{2,1}||$.

Proof. That $\mathcal{F}$ has envelope $||B_1||\,||H_{2,1}||$ is obvious, the envelope is square-integrable by (A1). That $\mathcal{F}$ is BUEI follows from Theorem 8.4 of Anthony and Bartlett (1999) which can be used to show that $\mathcal{F}$ is a VC class and hence satisfies the Uniform Entropy Integral condition. Finally, the measurability condition is met since any element in $\mathcal{F}$ can be arbitrarily closely approximated by an element in an analogous class whose parameters are restricted to be rational (seek Kosorok 2008). □

Lemma 3.6.8. *Assume (A1)-(A4). Then it follows that*

$$\mathbb{P}_n B_1^\intercal\left(\left[H_{2,1}^\intercal\mathbb{V}_n\right]_+ - \mathbb{U}_n\right)1_{T_n(H_{2,1})\le\lambda_n} = o_P(1).$$

Proof. As a first step, notice that using the definitions of $\mathbb{U}_n$ and $\mathbb{V}_n$ we have

$$\mathbb{P}_n B_1^\intercal \left(\left[H_{2,1}^\intercal \mathbb{V}_n \right]_+ - \mathbb{U}_n \right) 1_{T_n(H_{2,1}) \le \lambda_n} = \mathbb{P}_n B_1^\intercal \left(\left[H_{2,1}^\intercal \mathbb{V}_n \right]_+ - \mathbb{U}_n \right) 1_{T_n(H_{2,1}) \le \lambda_n} 1_{H_{2,1}^\intercal \beta_{2,1}^* \neq 0}.$$

The norm of the right hand side of the above display is, by appeal to Lemma (3.6.1) and the Cauchy-Schwartz inequality, bounded above by

$$2\mathbb{P}_n ||B_1||\, ||H_{2,1}|| 1_{T_n(H_{2,1}) \le \lambda_n} 1_{H_{2,1}^\intercal \beta_{2,1}^* \neq 0},$$

which we can subsequently write as

$$2P||B_1||\, ||H_{2,1}|| 1_{T_n(H_{2,1}) \le \lambda_n} 1_{H_{2,1}^\intercal \beta_{2,1}^* \neq 0} + 2(\mathbb{P}_n - P)||B_1||\, ||H_{2,1}|| 1_{T_n(H_{2,1}) \le \lambda_n} 1_{H_{2,1}^\intercal \beta_{2,1}^* \neq 0}. \tag{3.31}$$

The first term in (3.31) is $o_P(1)$ by Corollary (3.6.6). The second term in (3.31) is $o_P(1)$ by application of the uniform law of large numbers applied over the class of functions $\mathcal{F}$ defined in Lemma (3.6.7). □

Lemma 3.6.9. *Assume (A1)-(A4). Define* $\mathbb{V}_n^{(n)} \triangleq \sqrt{n}(\hat{\beta}_{2,1} - \delta_{2,1}^{(n)})$. *Then, under* $P_{n,\gamma}$*, it follows that*

$$\begin{aligned}
&\mathbb{P}_n B_1^\intercal \left(\left[H_{2,1}^\intercal \mathbb{V}_n^{(n)} \right]_+ - \sqrt{n} \left(\left[H_{2,1}^\intercal \hat{\beta}_{2,1} \right]_+ - \left[H_{2,1}^\intercal \delta_{2,1}^{(n)} \right]_+ \right) \right) 1_{T_n(H_{2,1}) \le \lambda_n} \\
&\quad = PB_1^\intercal \left(\left[H_{2,1}^\intercal \mathbb{V}_n^{(n)} \right]_+ - \left[H_{2,1}^\intercal (\mathbb{V}_n^{(n)} + \gamma) \right]_+ + \left[H_{2,1}^\intercal \gamma \right]_+ \right) 1_{H_{2,1}^\intercal \beta_{2,1}^* = 0} + o_{P_{n,\gamma}}(1).
\end{aligned}$$

Proof. Decompose the left hand side of the display in the statement of the lemma as

$$\begin{aligned}
&\mathbb{P}_n B_1^\intercal \left(\left[H_{2,1}^\intercal \mathbb{V}_n^{(n)} \right]_+ - \sqrt{n} \left(\left[H_{2,1}^\intercal \hat{\beta}_{2,1} \right]_+ - \left[H_{2,1}^\intercal \delta_{2,1}^{(n)} \right]_+ \right) \right) 1_{T_n(H_{2,1}) \le \lambda_n} 1_{H_{2,1}^\intercal \beta_{2,1}^* = 0} \\
&+ \mathbb{P}_n B_1^\intercal \left(\left[H_{2,1}^\intercal \mathbb{V}_n^{(n)} \right]_+ - \sqrt{n} \left(\left[H_{2,1}^\intercal \hat{\beta}_{2,1} \right]_+ - \left[H_{2,1}^\intercal \delta_{2,1}^{(n)} \right]_+ \right) \right) 1_{T_n(H_{2,1}) \le \lambda_n} 1_{H_{2,1}^\intercal \beta_{2,1}^* \neq 0}.
\end{aligned}$$

The norm of the second term in the above display is bounded above by

$$2\mathbb{P}_n||B_1||\,||H_{2,1}||1_{T_n(H_{2,1})\le\lambda_n}1_{H_{2,1}^\intercal\beta_{2,1}^*=0}||\mathbb{V}_n^{(n)}||$$

which we can subsequently decompose as

$$\begin{aligned}2P_{n,\gamma}||B_1||\,||H_{2,1}||1_{T_n(H_{2,1})\le\lambda_n}1_{H_{2,1}^\intercal\beta_{2,1}^*=0}||\mathbb{V}_n^{(n)}||&\\ +\,2(\mathbb{P}_n-P_{n,\gamma})||B_1||\,||H_{2,1}||1_{T_n(H_{2,1})\le\lambda_n}1_{H_{2,1}^\intercal\beta_{2,1}^*=0}||\mathbb{V}_n^{(n)}||.&\end{aligned}$$

It will be shown below that $\mathbb{V}_n^{(n)}=O_{P_{n,\gamma}}(1)$. Consequently, the contiguity of $P_{n,\gamma}$ and P paired with Corollary (3.6.6) and Slutskly's theorem show that the first term in the above display is $o_{P_{n,\gamma}}(1)$. Similarly, Lemma (3.6.7) combined with Glivenko-Cantelli results for contiguous alternatives (see Kosorok 2008) and Slutsky's theorem, ensure that the second term in the above display is also $o_{P_{n,\gamma}}(1)$.

We have shown that

$$\begin{aligned}&\mathbb{P}_nB_1^\intercal\left(\left[H_{2,1}^\intercal\mathbb{V}_n^{(n)}\right]_+-\sqrt{n}\left(\left[H_{2,1}^\intercal\hat{\beta}_{2,1}\right]_+-\left[H_{2,1}^\intercal\delta_{2,1}^{(n)}\right]_+\right)\right)1_{T_n(H_{2,1})\le\lambda_n}\\ &=\mathbb{P}_nB_1^\intercal\left(\left[H_{2,1}^\intercal\mathbb{V}_n^{(n)}\right]_+-\left[H_{2,1}^\intercal(\mathbb{V}_n^{(n)}+\gamma)\right]_++\left[H_{2,1}^\intercal\gamma\right]_+\right)1_{T_n(H_{2,1})\le\lambda_n}1_{H_{2,1}^\intercal\beta_{2,1}^*=0}+o_{P_{n,\gamma}}(1).\end{aligned}$$

Writing $1_{T_n(H_{2,1})\le\lambda_n}=1-1_{T_n(H_{2,1})>\lambda_n}$ and repeating the preceding argument allows us to remove omit the indicator $1_{T_n(H_{2,1})\le\lambda_n}$ from the above expression. Lastly, we need to argue that we can replace $\mathbb{P}_n$ with P in the above expression. Let $c\in\mathbb{R}^{\dim(B_1)}$ be arbitrary. Arguments given in the course of Lemma (3.6.4) can be used to show that the class of functions

$$\mathcal{F}\triangleq\left\{c^\intercal B_1^\intercal\left(\left[H_{2,1}^\intercal\delta\right]_+-\left[H_{2,1}^\intercal(\delta+\gamma)\right]_+-\left[H_{2,1}^\intercal\gamma\right]_+\right)1_{H_{2,1}^\intercal\beta_{2,1}^*=0}\,:\,\delta\in\mathcal{B}(M,0)\,,\,\gamma\in\mathbb{R}^{\dim(\beta_{2,1}^*)}\right\}$$

is a P-measurable BUEI class with square integrable envelope $2||B_1||\,||H_{2,1}||M$ and is thus Glivenko-Cantelli. Consequently, the quantity

$$(\mathbb{P}_n - P)B_1^\intercal \left(\left[H_{2,1}^\intercal \mathbb{V}_n^{(n)}\right]_+ - \left[H_{2,1}^\intercal(\mathbb{V}_n^{(n)} + \gamma)\right]_+ - \left[H_{2,1}^\intercal \gamma\right]_+ \right) 1_{H_{2,1}^\intercal \beta_{2,1}^* = 0}$$

converges to zero in P probability in $l^\infty(\mathbb{R}^{\dim(\beta_{2,1})^*})$, by contiguity, the same result follows in $P_{n,\gamma}$ probability. This proves the result. □

Lemma 3.6.10. *Assume (A1), (A2), and (A4). Let $\epsilon > 0$ and $K > 0$ be arbitrary. Fix a vector $c \in \mathbb{R}^{\dim(B_1)}$ and define, for each $\delta \in \mathcal{B}(K, 0) \subset \mathbb{R}^{\dim(\beta_{2,1}^*)}$, a function in $l^\infty(\mathbb{R}^{\dim(\beta_{2,1}^*)})$ given by*

$$g_\delta(\gamma) \triangleq Pc^\intercal B_1^\intercal \left(\left[H_{2,1}^\intercal(\delta + \gamma)\right]_+ - \left[H_{2,1}^\intercal \gamma\right]_+ - \left[H_{2,1}^\intercal \delta\right]_+ \right) 1_{H_{2,1}^\intercal \beta_{2,1}^* = 0}.$$

Then, there exists $M > 0$ so that

$$\sup_{\delta \in \mathcal{B}(K,0)} \left| \sup_{\gamma \in \mathcal{B}(M,0)} g_\delta(\gamma) - \sup_{\gamma \in \mathbb{R}^{\dim(\beta_{2,1}^*)}} g_\delta(\gamma) \right| < \epsilon.$$

Proof. The map from $\mathcal{B}(K, 0)$ into $l^\infty(\mathbb{R}^{\dim(\beta_{2,1}^*)})$ given by $\delta \mapsto g_\delta$ is uniformly continuous. To see this, note that for any $\delta_1, \delta_2 \in \mathcal{B}(K, 0)$ it follows that

$$||g_{\delta_1} - g_{\delta_2}||_{\mathbb{R}^{\dim(\beta_{2,1}^*)}} \leq 2P||c||\,||B_1||\,||H_{2,1}||\,||\delta_1 - \delta_2||.$$

Thus, we can choose be an $\epsilon/6$ net of points in the class $\{g_\delta\}_{\delta \in \mathcal{B}(K,0)}$, say $g_{\delta_1}, g_{\delta_2}, \ldots, g_{\delta_R}$. For a fixed value of δ, the function $g_\delta(\gamma)$ is uniformly continuous. For each $i =$

$1, 2, \ldots, R$ let M_i be such that

$$\sup_{\gamma \in \mathbb{R}^{\dim(\beta^*_{2,1})}} g_{\delta_i}(\gamma) - \sup_{\gamma \in \mathcal{B}(M_i,0} g_{\delta_i}(\gamma) \leq \epsilon/6,$$

and define $M^* \triangleq \max_i M_i$. Then, for any $M \geq M^*$ we have

$$\big| \sup_{\gamma \in \mathcal{B}(M,0)} g_\delta(\gamma) - \sup_{\gamma \in \mathbb{R}^{\dim(\beta^*_{2,1})}} g_\delta(\gamma) \big| \leq \inf_{1 \leq i \leq R} 2 \big|\big| g_{\delta_i} - g_\delta \big|\big|_{\mathbb{R}^{\dim(\beta^*_{2,1})}} + \epsilon/6 \leq \epsilon/2.$$

Furthermore, the above inequalities continue to hold if we talk a supremum over $\delta \in \mathcal{B}(K, 0)$, which proves the result. □

Lemma 3.6.11. *Assume (A3). Let $\mathcal{H}$ denote the hyperplane in $\mathbb{R}^{\dim(\beta^*_{2,1})}$ defined by the normal vector $\beta^*_{2,1}/||\beta^*_{2,1}||$. Let $\mathbb{P}_\mathcal{H}$ denote the projection matrix onto this space. Then, for any vector $v \in \mathbb{R}^{\dim(\beta^*_{2,1})}$, there exists a vector $\gamma \in \mathbb{R}^{\dim(\beta^*_{2,1})}$ such that $||\gamma - \sqrt{n}\beta^*_{2,1}||^2_\infty \leq \lambda_n$, and $\mathbb{P}_\mathcal{H}(v - \gamma) = 0$.*

Proof. Set $\gamma = v + \sqrt{n}\beta^*_{2,1}$. □

Corollary 3.6.12. *Assume (A1)-(A4). Let $M, K > 0$ be arbitrary and c be a fixed vector in $\mathbb{R}^{\dim(\beta^*_{2,1})}$. Let g_δ be defined as in Lemma (3.6.10). Define $\mathcal{S}_n$ to be the set given by $\{\gamma \in \mathbb{R}^{\dim(\beta^*_{2,1})} : ||\gamma - \sqrt{n}\beta^*_{2,1}||_\infty \leq \lambda_n\}$. Then, for sufficiently large n it follows that*

$$\inf_{\delta \in \mathcal{B}(K,0)} \left(\sup_{\gamma \in \mathcal{S}_n} g_\delta(\gamma) - \sup_{\gamma \in \mathcal{B}(M,0)} g_\delta(\gamma) \right) \geq 0.$$

Proof. Let $\mathcal{H}$ and $\mathbb{P}_\mathcal{H}$ be as defined in the preceding lemma. Recall that for any pair of vectors $h \in \mathcal{H}$ and $v \in \mathbb{R}^{\dim(\beta^*_{2,1})}$ it follows that $v^\intercal h = (\mathbb{P}_\mathcal{H} v)^\intercal h$. The form of g_δ shows that $g_\delta(\gamma) = g_\delta(\mathbb{P}_\mathcal{H}\gamma)$. Choosing n sufficiently large so that $M < \lambda_n$, implies, by appeal to the preceding Lemma, that the projection of $\mathcal{B}(M, 0)$ onto $\mathcal{H}$ is contained in the projection of $\mathcal{S}_n$ onto $\mathcal{H}$. This proves the result. □

Corollary 3.6.13. *Assume (A1)-(A4). Define* $\mathcal{S}_n \triangleq \left\{\delta \in \mathbb{R}^{\dim(\beta^*_{2,1})} : ||\delta - \sqrt{n}\beta^*_{2,1}||^2_\infty \leq \lambda_n\right\}$, *and let* $c \in \mathbb{R}^{\dim(B_1)}$ *be arbitrary. Let* $\epsilon > 0$ *be arbitrary and let* g_δ *be as defined in Lemma (3.6.10). Then,*

$$P\left(\sup_{\gamma\in\mathbb{R}^{\dim(\beta^*_{2,1})}} g_{\mathbb{V}_n}(\gamma) - \sup_{\gamma\in\mathcal{S}_n} g_{\mathbb{V}_n}(\gamma) > \epsilon\right) \leq \eta,$$

for sufficiently large n.

Proof. Choose K sufficiently large so that the probability that $||\mathbb{V}_n|| > K$ is less than $\eta/2$. Then, the probability in the statement of this Corollary is bounded above by

$$\eta/2 + 1_{\sup_{\delta\in\mathcal{B}(K,0)}\left[\sup_{\gamma\in\mathbb{R}^{\dim(\beta^*_{2,1})}} g_\delta(\gamma) - \sup_{\gamma\in\mathcal{S}_n} g_\delta(\gamma)\right] > \epsilon}.$$

As a next step, write

$$\left[\sup_{\gamma\in\mathbb{R}^{\dim(\beta^*_{2,1})}} g_\delta(\gamma) - \sup_{\gamma\in\mathcal{S}_n} g_\delta(\gamma)\right] = \left[\sup_{\gamma\in\mathbb{R}^{\dim(\beta^*_{2,1})}} g_\delta(\gamma) - \sup_{\gamma\in\mathcal{B}(M,0)} g_\delta(\gamma)\right] + \left[\sup_{\gamma\in\mathcal{B}(M,0)} g_\delta(\gamma) - \sup_{\gamma\in\mathcal{S}_n} g_\delta(\gamma)\right]$$

By appeal to Lemma (3.6.10) we can choose M sufficiently large so that

$$\sup_{\delta\in\mathcal{B}(K,0)}\left[\sup_{\gamma\in\mathbb{R}^{\dim(\beta^*_{2,1})}} g_\delta(\gamma) - \sup_{\gamma\in\mathcal{B}(M,0)} g_\delta(\gamma)\right] \leq \epsilon/2.$$

Furthermore, by appeal to Corollary (3.6.13) and the proof of that result it follows that for sufficiently large n

$$\mathbb{P}_{\mathcal{H}}\mathcal{B}(M,0) \triangleq \{\mathbb{P}_{\mathcal{H}}v : v \in \mathcal{B}(M,0)\} \subseteq \mathbb{P}_{\mathcal{H}}\mathcal{S}_n \triangleq \{\mathbb{P}_{\mathcal{H}}\gamma : \gamma \in \mathcal{S}_n\},$$

where $\mathbb{P}_{\mathcal{H}}$ is as defined in Lemma (3.6.11). Recall $g_\delta(\gamma) = g_{\delta(}(\mathbb{P}_{\mathcal{H}}\gamma)$, the result is proved by noting that as a consequence of the preceding discussion we have

$$\left[\sup_{\gamma\in\mathcal{B}(M,0)} g_\delta(\gamma) - \sup_{\gamma\in\mathcal{S}_n} g_\delta(\gamma)\right] \leq 0.$$

□

The preceding series of results show that, without a loss of generality, the supremum in the definition of $\mathcal{U}(c)$ can be viewed as being either unrestricted or as being taken over a large ball centered at the origin. In addition, notice that the bootstrap analogue of the set $\mathcal{S}_n$, say, $\mathcal{S}_n^{(b)}$ is given by

$$\mathcal{S}_n^{(b)} \triangleq \left|\left|\gamma - \sqrt{n}\hat{\beta}_{2,1}\right|\right|_\infty^2 \leq \lambda_n$$

which is equivalent to

$$\left|\left|\gamma - \mathbb{V}_n - \sqrt{n}\beta_{2,1}^*\right|\right| \leq \lambda_n.$$

and thus, using the fact that $\mathbb{V}_n = O_P(1)$, bootstrap analogues of the preceding series of results hold with probability tending to one. That is, we can regard the supremum in the definition of $\mathcal{U}^{(b)}(c)$ as being unrestricted or having been taken over a suitably large ball centered at the origin.

Proof. Theorem 2.1, part 3. Notice that $\mathcal{U}(c) - c^\intercal\sqrt{n}(\hat{\beta}_1 - \beta_1^*)$ is equal to

$$\begin{aligned}&c^\intercal\Sigma_{1,n}^{-1}\mathbb{P}_n B_1^\intercal\left(\left[H_{2,1}^\intercal\mathbb{V}_n\right]_+ - \mathbb{U}_n\right)1_{T_n(H_{2,1})\leq\lambda_n}\\ &+\sup_{\gamma\in\mathbb{R}^{\dim(\beta_{2,1}^*)}} c^\intercal\Sigma_{1,n}^{-1}\mathbb{P}_n B_1^\intercal\left(\left[H_{2,1}^\intercal(\mathbb{V}_n+\gamma)\right]_+ - \left[H_{2,1}^\intercal\gamma\right]_+ - \left[H_{2,1}^\intercal\mathbb{V}_n\right]_+\right)1_{\frac{\left(H_{2,1}^\intercal(\mathbb{V}_n+\gamma)\right)^2}{H_{2,1}^\intercal\Sigma_{2,n}^{(2,2)}H_{2,1}}\leq\lambda_n}1_{T_n(H_{2,1})\leq\lambda_n},\end{aligned}$$

the first term of which is $o_P(1)$ by Lemma (3.6.8). An argument similar to the one given in the proof of part 1 of the theorem shows that the preceding expression is equal to

$$\sup_{\gamma \in \mathbb{R}^{\dim(\beta^*_{2,1})}} c^\intercal \Sigma^{-1}_{1,\infty} P B_1^\intercal \left(\left[H_{2,1}^\intercal(\mathbb{V}_n + \gamma)\right]_+ - \left[H_{2,1}^\intercal \gamma\right]_+ - \left[H_{2,1}^\intercal \mathbb{V}_n\right]_+ \right) 1_{H_{2,1}^\intercal \beta^*_{2,1}=0} + o_P(1).$$

The desired result follows from the continuous mapping theorem. To see this, it suffices to show that the map from $\mathbb{R}^{\dim(\beta^*_{2,1})}$ into $l^\infty(\mathbb{R}^{\dim(\beta^*_{2,1})})$ defined by $\delta \mapsto f_\delta(\gamma)$ where

$$f_\delta(\gamma) \triangleq c^\intercal \Sigma^{-1}_{1,\infty} P B_1^\intercal \left(\left[H_{2,1}^\intercal(\delta + \gamma)\right]_+ - \left[H_{2,1}^\intercal \gamma\right]_+ - \left[H_{2,1}^\intercal \delta\right]_+ \right) 1_{H_{2,1}^\intercal \beta^*_{2,1}=0}$$

is continuous. That is, given arbitrary $\epsilon > 0$ we must show there exists $\eta > 0$ so that $||\delta - \Lambda|| < \eta$ implies $\sup_{\gamma \in \mathbb{R}^{\dim(\beta^*_{2,1})}} \left|f_\delta(\gamma) - f_\Lambda(\gamma)\right| \leq \epsilon$. Notice that

$$\left|f_\delta(\gamma) - f_\Lambda(\gamma)\right| \leq \sup_{\gamma \in \mathbb{R}^{\dim(\beta^*_{2,1})}} 2\, P\left|c^\intercal \Sigma^{-1}_{1,\infty} B_1^\intercal\right| \left| \left[H_{2,1}^\intercal(\delta + \gamma)\right]_+ - \left[H_{2,1}^\intercal(\Lambda + \gamma)\right]_+ \right| 1_{H_{2,1}^\intercal \beta^*_{2,1}=0}$$

which, by the Cauchy-Schwartz inequality, is further bounded above by

$$||\delta - \Lambda||\, P\left|c^\intercal \Sigma^{-1}_{1,\infty} B_1^\intercal\right| ||H_{2,1}||\, 1_{H_{2,1}^\intercal \beta^*_{2,1}=0}.$$

The above term can be made arbitrarily small which proves the continuity.

We now derive the limiting distribution of $\mathcal{U}(c)$ under local alternatives of the form described in (A4). Utilizing Lemma (3.6.9) it follows that $\mathcal{U}(c) - c^\intercal \sqrt{n}(\hat{\beta}_1 - \delta_1^{(n)})$ is equal to

$$\begin{aligned} &c^\intercal \Sigma^{-1}_{1,n} \mathbb{P}_n B_1^\intercal \left(\left[H_{2,1}^\intercal \mathbb{V}_n^{(n)}\right]_+ - \left[H_{2,1}^\intercal(\mathbb{V}_n^{(n)} + \gamma)\right]_+ + \left[H_{2,1}^\intercal \gamma\right]_+ \right) 1_{H_{2,1}^\intercal \beta^*_{2,1}=0} \\ &+ \sup_{\rho \in \mathbb{R}^{\dim(\beta^*_{2,1})}} c^\intercal \Sigma^{-1}_{1,n} \mathbb{P}_n B_1^\intercal \left(\left[H_{2,1}^\intercal \left(\mathbb{V}_n^{(n)} + \rho + \gamma\right)\right]_+ - \left[H_{2,1}^\intercal(\rho + \gamma)\right]_+ - \left[H_{2,1}^\intercal \mathbb{V}_n^{(n)}\right]_+ \right) 1_{T_n(H_{2,1}) \leq \lambda_n} \end{aligned}$$

$$+ o_{P_{n,\gamma}}(1).$$

where $\mathbb{V}_n^{(n)} \triangleq \sqrt{n}(\hat{\beta}_{2,1} - \delta_{2,1}^{(n)})$. Since the supremum is unrestricted, we can replace ρ with $\rho - \gamma$ to obtain

$$c^\intercal \Sigma_{1,n}^{-1} \mathbb{P}_n B_1^\intercal \left(\left[H_{2,1}^\intercal \mathbb{V}_n^{(n)} \right]_+ - \left[H_{2,1}^\intercal (\mathbb{V}_n^{(n)} + \gamma) \right]_+ + \left[H_{2,1}^\intercal \gamma \right]_+ \right) 1_{H_{2,1}^\intercal \beta_{2,1}^* = 0}$$
$$\sup_{\rho \in \mathbb{R}^{\dim(\beta_{2,1}^*)}} c^\intercal \Sigma_{1,n}^{-1} \mathbb{P}_n B_1^\intercal \left(\left[H_{2,1}^\intercal \left(\mathbb{V}_n^{(n)} + \rho \right) \right]_+ - \left[H_{2,1}^\intercal \rho \right]_+ - \left[H_{2,1}^\intercal \mathbb{V}_n^{(n)} \right]_+ \right) 1_{T_n(H_{2,1}) \le \lambda_n}.$$

It was established in the proof of part 2 of Theorem 2.1. that $\mathbb{V}_n^{(n)}$ has the same limiting distribution under $P_{n,\gamma}$ as $\sqrt{n}(\hat{\beta}_{2,1} - \beta_{2,1}^*)$ under P. Consequently, the proof for the fixed alternatives case can be applied to finish the remainder of this proof. □

The proof of Theorem 2.2. is made simpler by first introducing the following lemmas.

Lemma 3.6.14. *Let $X_n^{(b)}$ be a bootstrap process in $\mathbb{R}^p$ which converges in distribution to tight, continuous, random variable X in probability (as defined in the main body of the paper). Then, for arbitrary $\delta > 0$ there exists M sufficiently large so that*

$$P_{\mathcal{M}} \left(||X_n^{(b)}|| > M \right) \le \delta + r_n$$

where r_n is $o_P(1)$ and r_n does not depend on either M or δ.

Proof. The bootstrap continuous mapping theorem (see proposition 10.7 Kosorok 2008) implies that $||X_n^{(b)}||$ converges in distribution to $||X||$ in probability. Since X is tight, we can choose M sufficiently large so that $P(||X|| > M) \le \delta$. Define,

$$r_n \triangleq \sup_{t \in \mathbb{R}} \left| P_{\mathcal{M}}(||X_n^{(b)}|| \le t) - P(||X|| \le t) \right|$$

then by Lemma 10.10 in Kosorok (2008) r_n is $o_P(1)$. Since r_n does not depend on either M or δ and

$$P_M(||X_n^{(b)}|| > M) \leq P(||X|| > M) + r_n \leq \delta + r_n,$$

the result is proved. □

Lemma 3.6.15. *Assume (A1)-(A3). Then for any* $\epsilon > 0$*,*

$$P_{\mathcal{M}}\left(\hat{\mathbb{P}}_n^{(b)}||B_1||\,||H_{2,1}||1_{(H_{2,1}^{\intercal}\hat{\beta}_{2,1}^{(b)})(H_{2,1}^{\intercal}\beta_{2,1}^{*})<0} > \epsilon\right)$$

converges to zero (outer) almost surely.

Proof. This proof proceeds much in the same way as Lemma (3.6.3). As a first step we decompose $\hat{\mathbb{P}}_n^{(b)}||B_1||\,||H_{2,1}||1_{(H_{2,1}^{\intercal}\hat{\beta}_{2,1}^{(b)})(H_{2,1}^{\intercal}\beta_{2,1}^{*})<0}$ into the following two parts

$$\hat{\mathbb{P}}_n^{(b)}||B_1||\,||H_{2,1}||1_{H_{2,1}^{\intercal}\hat{\beta}_{2,1}^{(b)}<0}1_{H_{2,1}^{\intercal}\beta_{2,1}^{*}>0} + \hat{\mathbb{P}}_n^{(b)}||B_1||\,||H_{2,1}||1_{H_{2,1}^{\intercal}\hat{\beta}_{2,1}^{(b)}>0}1_{H_{2,1}^{\intercal}\beta_{2,1}^{*}<0}. \quad (3.32)$$

From the form of preceding decomposition we see that it suffices to show that

$$P_{\mathcal{M}}\left(\hat{\mathbb{P}}_n^{(b)}||B_1||\,||H_{2,1}||1_{H_{2,1}^{\intercal}\hat{\beta}_{2,1}^{(b)}<0}1_{H_{2,1}^{\intercal}\beta_{2,1}^{*}>0} > \epsilon/2\right)$$

converges outer almost surely to zero since we can repeat the argument with $-H_{2,1}$ in place of $H_{2,1}$.

The argument and result of Lemma (3.6.7) show that the class

$$\mathcal{F} \triangleq \left\{||B_1||\,||H_{2,1}||1_{H_{2,1}^{\intercal}\delta<0}1_{H_{2,1}^{\intercal}\beta_{2,1}^{*}>0}\,,\ \delta \in \mathbb{R}^{\dim(\beta_{2,1}^{*})}\right\}$$

is a P-measurable VC class and hence strong Glivenko-Cantelli (Kosorok 2008; see also Corollary 3.12 of van der Geer 2000). Thus,

$$\hat{\mathbb{P}}_n^{(b)}||B_1||\,||H_{2,1}||1_{H_{2,1}^\intercal\hat{\beta}_{2,1}^{(b)}<0}1_{H_{2,1}^\intercal\beta_{2,1}^*>0} \leq P||B_1||\,||H_{2,1}||1_{H_{2,1}^\intercal\hat{\beta}_{2,1}^{(b)}<0}1_{H_{2,1}^\intercal\beta_{2,1}^*>0}+||\hat{\mathbb{P}}_n^{(b)}-P||_{\mathcal{F}}^*$$

where $*$ denotes a measurable majorant with respect to both the bootstrap weights and the observed data. Since $\mathcal{F}$ is measurable and strong Glivenko-Cantelli, it follows that $P_{\mathcal{M}}\left(||\hat{\mathbb{P}}_n^{(b)} - P||_{\mathcal{F}}^* > \epsilon/6\right)$ converges to zero outer almost surely (Lemma 3.6.16 van der Vaart and Wellner 1996). Next, we consider the term $P||B_1||\,||H_{2,1}||1_{H_{2,1}^\intercal\hat{\beta}_{2,1}^{(b)}<0}1_{H_{2,1}^\intercal\beta_{2,1}^*>0}$ which we can bound above by

$$P||B_1||\,||H_{2,1}||1_{H_{2,1}^\intercal\hat{\beta}_{2,1}^{(b)}<0}1_{\frac{H_{2,1}^\intercal\beta_{2,1}^*}{||H_{2,1}||}>\eta} + P||B_1||\,||H_{2,1}||1_{\frac{H_{2,1}^\intercal\beta_{2,1}^*}{||H_{2,1}||}\leq\eta},$$

where $\eta > 0$ is arbitrary. The second term in the above decomposition can be made less than $\epsilon/6$ by sufficiently small choice of η. Let η be fixed at this value. Notice that the event

$$\left\{H_{2,1}^\intercal\hat{\beta}_{2,1}^{(b)} < 0, \frac{H_{2,1}^\intercal\beta_{2,1}^*}{||H_{2,1}||} > \eta\right\}$$

is contained in the event

$$\left\{\frac{H_{2,1}^\intercal(\mathbb{V}_n^{(b)} + \mathbb{V}_n)}{||H_{2,1}||} < -\frac{\sqrt{n}H_{2,1}^\intercal\beta_{2,1}^*}{||H_{2,1}||}, \frac{\sqrt{n}H_{2,1}^\intercal\beta_{2,1}^*}{||H_{2,1}||} > \eta\right\},$$

where $\mathbb{V}_n^{(b)} \triangleq \sqrt{n}(\hat{\beta}_{2,1}^{(b)} - \hat{\beta}_{2,1})$. We can further contain the event in the preceding display in the event

$$\left\{||\mathbb{V}_n^{(b)} + \mathbb{V}_n|| > \sqrt{n}\eta\right\}.$$

Thus, we have shown that

$$P||B_1||\,||H_{2,1}||1_{H_{2,1}^{\intercal}\hat{\beta}_{2,1}^{(b)}<0}1_{\frac{H_{2,1}^{\intercal}\beta_{2,1}^{*}}{||H_{2,1}||}>\eta} \leq 1_{||\mathbb{V}_n^{(b)}+\mathbb{V}_n||>\sqrt{n}\eta}P||B_1||\,||H_{2,1}||.$$

So that,

$$P_{\mathcal{M}}\left(P||B_1||\,||H_{2,1}||1_{H_{2,1}^{\intercal}\hat{\beta}_{2,1}^{(b)}<0}1_{\frac{H_{2,1}^{\intercal}\beta_{2,1}^{*}}{||H_{2,1}||}>\eta} > \epsilon/6\right) \leq P_{\mathcal{M}}\left(1_{||\mathbb{V}_n^{(b)}+\mathbb{V}_n||>\sqrt{n}\eta}P||B_1||\,||H_{2,1}|| > \epsilon/6\right),$$

the right hand side of which is less than $P_{\mathcal{M}}\left(||\mathbb{V}_n^{(b)} + \mathbb{V}_n|| > \sqrt{n}\eta\right)$, which we can write as

$$P_{\mathcal{M}}\left(\frac{||\mathbb{V}_n^{(b)} + \mathbb{V}_n||}{\sqrt{log\ \log n}} > \sqrt{\frac{n}{\log\log n}}\eta\right) \leq P_{\mathcal{M}}\left(\frac{||\mathbb{V}_n^{(b)}||}{\sqrt{\log\log n}} > \sqrt{\frac{n}{\log\log n}}\eta - \frac{||\mathbb{V}_n||}{\sqrt{\log\log n}}\right),$$

where the last inequality follows from the triangle inequality. The right hand side of the above display converges to zero almost surely since $||\mathbb{V}_n||/\sqrt{\log\log n}$ is bounded almost surely by the LIL, and from Lemma (3.6.14) it follows that $||\mathbb{V}_n^{(b)}||$ is $O_{P_{\mathcal{M}}}(1)$. The result follows. □

Lemma 3.6.16 (Bootstrap Consistent Treatment Assignment)**.** *Assume (A1)-(A4). Define* $T_n(h_{2,1})^{(b)} \triangleq \frac{n(h_{2,1}^{\intercal}\hat{\beta}_{2,1}^{(b)})^2}{h_{2,1}^{\intercal}(\Sigma_{2,n}^{(2,2)})^{(b)}h_{2,1}}$ *to be the bootstrap analog of* $T_n(h_{2,1})$*. Then, the following results hold.*

1. $P_{\mathcal{M}}\left(\sup_{h_{2,1}} 1_{T_n^{(b)}(h_{2,1})>\lambda_n}1_{h_{2,1}^{\intercal}\beta_{2,1}^{*}=0} = 1\right)$ *converges to zero in probability.*

2. *For any* $\epsilon > 0$*, there exists a set* B_ϵ *so that* $P\left(H_{2,1} \in B_\epsilon\right) \geq 1 - \epsilon$ *and*

$$P_M\left(\sup_{h_{2,1}\in B_\epsilon} 1_{T_n^{(b)}(h_{2,1})\leq\lambda_n}1_{h_{2,1}^{\intercal}\beta_{2,1}^{*}\neq 0} = 1\right)$$

converges to zero in probability.

Proof. We begin with the proof of part 1. Let $h_{2,1}$ be fixed non-zero vector satisfying $h_{2,1}^{\intercal}\beta_{2,1}^{*}=0$, then

$$\frac{T_n^{(b)}(h_{2,1})}{\lambda_n} = \frac{\left(h_{2,1}^{\intercal}\left(\mathbb{V}_n^{(b)}/\sqrt{\lambda_n}+\mathbb{V}_n/\sqrt{\lambda_n}\right)\right)^2}{h_{2,1}^{\intercal}\Sigma_{2,n}^{(2,2)(b)}h_{2,1}} \leq \frac{\left(||\mathbb{V}_n^{(b)}||_{\infty}/\sqrt{\lambda_n}+||\mathbb{V}_n||_{\infty}/\sqrt{\lambda_n}\right)^2}{\sigma_{2,n}^{(2,2)(b)}},$$

where $||.||_{\infty}$ denotes the usual sup norm $z \mapsto \max_i |z_i|$ and $\sigma_{2,n}^{(2,2)(b)}$ denotes the smallest eigenvalue of $\Sigma_{2,n}^{(2,2)(b)}$. For any $\epsilon > 0$ we can bound the term on the right hand side of the above display by

$$\frac{\left(||\mathbb{V}_n^{(b)}||_{\infty}/\sqrt{\lambda_n}+\epsilon\right)^2}{\sigma_{2,n}^{(2,2)(b)}} + \frac{\left(||\mathbb{V}_n^{(b)}||_{\infty}/\sqrt{\lambda_n}+||\mathbb{V}_n||_{\infty}/\sqrt{\lambda_n}\right)^2}{\sigma_{2,n}^{(2,2)(b)}} 1_{\frac{||\mathbb{V}_n||_{\infty}}{\sqrt{\lambda_n}}>\epsilon}.$$

Observe that the above bound does not depend on $h_{2,1}$. Thus, we have

$$P_{\mathcal{M}}\left(\sup_{h_{2,1}:h_{2,1}^{\intercal}\beta_{2,1}^{*}=0} T_n^{(b)}(h_{2,1}) > \lambda_n\right) \leq P_{\mathcal{M}}\left(||\mathbb{V}_n^{(b)}||_{\infty}/\sqrt{\lambda_n} > \sqrt{\sigma_{2,n}} - \epsilon\right) + 1_{\frac{||\mathbb{V}_n||_{\infty}}{\sqrt{\lambda_n}}>\epsilon}. \tag{3.33}$$

Let $\sigma_{2,\infty}^{(2,2)}$ denote the smallest eigenvalue of $\Sigma_{2,\infty}^{(2,2)}$. If we choose $0 < \epsilon << \sigma_{2,\infty}^{(2,2)}$ then the term on the right hand side of (3.33) is seen, by appeal to the bootstrap LLN (see Athreya 1983; Csorgo and Rosalsky 2003), to be $o_P(1)$. If $h_{2,1}^{\intercal}\beta_{2,1}^{*} \neq 0$ then clearly $1_{T_n(h_{2,1})>\lambda_n}1_{h_{2,1}^{\intercal}\beta_{2,1}^{*}=0} \equiv 0$. We have shown that $P_{\mathcal{M}}(\sup_{h_{2,1}} 1_{T_n(h_{2,1})>\lambda_n}1_{h_{2,1}^{\intercal}\beta_{2,1}^{*}=0} = 1)$ converges to zero in probability.

The proof of part 2 of the lemma follows a similar line of argument. For any $\epsilon > 0$ we can choose $\delta > 0$ so that

$$P\left(0 < \frac{|H_{2,1}^{\intercal}\beta_{2,1}^{*}|}{||H_{2,1}||} \leq \delta\right) \leq \epsilon.$$

Define B_ϵ to be the event $\left\{H_{2,1}^{\intercal}\beta_{2,1}^{*}=0\right\}\bigcup\left\{\frac{|H_{2,1}^{\intercal}\beta_{2,1}^{*}|}{||H_{2,1}||} > \delta\right\}$ so that $P(H_{2,1} \in B_\epsilon) \geq$

$1-\epsilon$. Let $h_{2,1}$ be any value in B_ϵ satisfying $h_{2,1}^{\intercal}\beta_{2,1}^{*} \neq 0$, then

$$\frac{T_n(h_{2,1})}{\lambda_n} = \frac{\left(h_{2,1}^{\intercal}\left(\mathbb{V}_n^{(b)}/\sqrt{\lambda_n} + \mathbb{V}_n/\sqrt{\lambda_n} + \sqrt{\tfrac{n}{\lambda_n}}\beta_{2,1}^{*}\right)\right)^2}{h_{2,1}^{\intercal}\Sigma_{2,n}^{(2,2)(b)}h_{2,1}}.$$

The right hand side of the above display is bounded below by

$$\frac{\left(h_{2,1}^{\intercal}\left(\mathbb{V}_n^{(b)}/\sqrt{\lambda_n} + \mathbb{V}_n/\sqrt{\lambda_n} + \sqrt{\tfrac{n}{\lambda_n}}\beta_{2,1}^{*}\right)\right)^2}{||h_{2,1}||^2\nu_{2,n}^{(2,2)(b)}}$$

where $\nu_{2,n}^{(2,2)(b)}$ denotes the largest eigenvalue of $\Sigma_{2,n}^{(2,2)(b)}$. Using the triangle inequality and the fact that $h_{2,1}$ is in B_ϵ and $h_{2,1}^{\intercal}\beta_{2,1}^{*} \neq 0$, we can obtain an even smaller lower bound of

$$\left[\frac{\sqrt{\tfrac{n}{\lambda_n}}\sqrt{\delta} - ||\mathbb{V}_n^{(b)}||_\infty/\sqrt{\lambda_n} - ||\mathbb{V}_n||_\infty/\sqrt{\lambda_n}}{\sqrt{\nu_{2,n}^{(2,2)(b)}}}\right]_+^2.$$

The preceding sequence of bounds show that

$$P_{\mathcal{M}}\left(\sup_{h_{2,1}\in B_\epsilon} 1_{T_n^{(b)}(h_{2,1})\leq\lambda_n} 1_{h_{2,1}^{\intercal}\beta_{2,1}^{*}\neq 0} = 1\right) \leq P_{\mathcal{M}}\left(\left[\frac{\sqrt{\tfrac{n}{\lambda_n}}\sqrt{\delta} - ||\mathbb{V}_n^{(b)}||_\infty/\sqrt{\lambda_n} - ||\mathbb{V}_n||_\infty/\sqrt{\lambda_n}}{\sqrt{\nu_{2,n}^{(2,2)(b)}}}\right]_+ \leq 1\right)$$

the right hand side of which is further bounded above by

$$P_{\mathcal{M}}\left(\sqrt{\delta}\sqrt{\frac{n}{\lambda_n}} - M - \sqrt{\nu_{2,n}^{(2,2)(b)}} \leq ||\mathbb{V}_n^{(b)}||_\infty/\sqrt{\lambda_n}\right) + 1_{\frac{||\mathbb{V}_n||_\infty}{\sqrt{\lambda_n}}>M}$$

for any $M > 0$. The preceding quantity is $o_P(1)$ by appeal to the bootstrap LLN (Athreya 1983; Csorgo and Rosalsky 2003) applied to $\nu_{2,n}^{(2,2)(b)}$ and Lemma (3.6.14) applied to $\mathbb{V}_n^{(b)}$. □

Corollary 3.6.17. *Assume (A1)-(A4). Let $\epsilon > 0$ be arbitrary. Then, it follows that* *both of the quantities:*

1. $P_{\mathcal{M}}\left(P||B_1||\,||H_{2,1}||1_{T_n^{(b)}(H_{2,1})\leq\lambda_n}1_{H_{2,1}^{\intercal}\beta_{2,1}^*\neq 0} > \epsilon\right)$,
2. $P_{\mathcal{M}}\left(P||B_1||\,||H_{2,1}||1_{T_n^{(b)}(H_{2,1})>\lambda_n}1_{H_{2,1}^{\intercal}\beta_{2,1}^*=0} > \epsilon\right)$,

converge to zero in probability.

Proof. To prove part 1 we first observe that

$$P||B_1||\,||H_{2,1}||1_{T_n^{(b)}(H_{2,1})\leq\lambda_n}1_{H_{2,1}^{\intercal}\beta_{2,1}^*\neq 0} \leq P||B_1||\,||H_{2,1}||1_{H_{2,1}\in B_\eta}1_{T_n^{(b)}(H_{2,1})\leq\lambda_n}1_{H_{2,1}^{\intercal}\beta_{2,1}^*\neq 0} + P||B_1||\,||H_{2,1}||1_{H_{2,1}\notin B_\eta},$$

where B_η was defined in (3.6.16). We can choose η sufficiently small so that the second term on the right hand side of the above display is below $\epsilon/2$. The first term on the right hand side of the above display is bounded above by

$$\sup_{h_{2,1}\in B_\eta} 1_{T_n^{(b)}(h_{2,1})\leq\lambda_n}1_{h_{2,1}^{\intercal}\beta_{2,1}^*\neq 0}\left(P||B_1||\,||H_{2,1}||\right).$$

Using Lemma (3.6.16) it follows that

$$P_{\mathcal{M}}\left(\sup_{h_{2,1}\in B_\eta} 1_{T_n^{(b)}(h_{2,1})\leq\lambda_n}1_{h_{2,1}^{\intercal}\beta_{2,1}^*\neq 0}\left(P||B_1||\,||H_{2,1}||\right) > \epsilon\right)$$

converges to zero in probability. The proof of the second part of the theorem follows from an identical argument replacing B_η with all of $\mathbb{R}^{\dim(\beta_{2,1}^*)}$. □

Lemma 3.6.18. *Assume (A1)-(A4). Then, it follows that*

$$\sqrt{n}\hat{\mathbb{P}}_n^{(b)}B_1^{\intercal}\left(\left[H_{2,1}^{\intercal}\hat{\beta}_{2,1}^{(b)}\right]_+ - \left[H_{2,1}^{\intercal}\hat{\beta}_{2,1}\right]_+\right)1_{H_{2,1}^{\intercal}\beta_{2,1}^*\neq 0}$$

converges in distribution to $PB_1^{\intercal}H_{2,1}^{\intercal}\mathbb{V}_\infty 1_{H_{2,1}^{\intercal}\beta_{2,1}^*>0}$ *in probability.*

Proof. □

Lemma 3.6.19 (Computational Equivalence Lemma). *Let $\omega_1, \omega_2, \ldots, \omega_r$ be arbitrary negative weights, and $\omega_{r+1}, \omega_{r+2}, \ldots, \omega_M$ be arbitrary positive weights. Let $a_1, a_2, \ldots, a_M$ and γ be a fixed vectors in $\mathbb{R}^{\dim(\beta^*_{2,1})}$. Then,*

$$\arg \sup_{\delta \in \mathbb{R}^{\dim(\beta^*_{2,1})}} \sum_{i=1}^{M} \omega_i \left([a_i^\intercal (\gamma + \delta)]_+ - [a_i^\intercal \delta]_+ \right)$$

is equal to

$$\arg \inf_{\delta \in \mathbb{R}^{\dim(\beta^*_{2,1})}} \Bigg(\sum_{i=1}^{r} |\omega_i| \min\left(|a_i^\intercal \gamma|, |a_i^\intercal(\gamma+\delta)|\right) 1_{(a_i^\intercal \gamma)(a_i^\intercal(\gamma+\delta)>0} \\ + \sum_{i=r+1}^{M} \omega_i \min\left(|a_i^\intercal \gamma|, |a_i^\intercal \delta|\right) 1_{(a_i^\intercal \gamma)(a_i^\intercal \delta)<0} \Bigg).$$

Proof. Note that $\sup_{b\in\mathbb{R}} \left([a_i^\intercal \gamma + b]_+ - [b]_+\right) = [a_i^\intercal \gamma]_+$ and that this supremum is attained whenever $(a_i^\intercal \gamma) b \geq 0$. The regret incurred at a value b for which $(a_i^\intercal \gamma) b < 0$ is $[a_i^\intercal \gamma]_+ - \left([a_i^\intercal \gamma + b]_+ - [b]_+\right) = \min(|a_i^\intercal \gamma|, |b|)$. Similarly, $\inf_{b\in\mathbb{R}} \left([a_i^\intercal \gamma + b]_+ - [b]_+\right) = [a_i^\intercal \gamma]_-$, where $[.]_-$ is the map $z \mapsto \min(z, 0)$. The preceding inf is attained whenever $(a_i^\intercal \gamma + b) a_i^\intercal \gamma \leq 0$. The regret incurred at a value b for which $(a_i^\intercal \gamma + b) a_i^\intercal \gamma > 0$ is $\left([a_i^\intercal \gamma + b]_+ - [b]_+\right) - [a_i^\intercal \gamma]_- = \min(|a_i^\intercal \gamma|, |a_i^\intercal \gamma + b|)$. Since maximizing an objective is equivalent to minimizing the regret the result is proved. □

The preceding lemma shows that computing the ACI is computationally equivalent to Ψ-learning. This is somewhat of a bittersweet result since on the one hand its is known Ψ-learning is non-convex and cannot be solved using usual out-of-the-box solvers, on the other hand, substantial efforts have been made to produce efficient algorithms yielding (nearly) exact solutions (Lui 2004; Lui 2006).

Proof of Theorem 2.2. The proof of this result requires a bit of bookkeeping. It will be convenient to divide it into separate steps. Throughout, we will let $\mathbb{G}_n^{(b)}$ denote $\sqrt{n}(\hat{\mathbb{P}}_n^{(b)} - \mathbb{P}_n)$. Furthermore, we will make frequent use of the fact that under (A1)-(A3) $\sqrt{n}(\hat{\beta}_2^{(b)} - \beta_2^*)$ converges weakly in probability to $\mathbb{V}_\infty$. This result follows from the bootstrap central limit theorem (see Bickel and Freedman 1981; Csorgo and Mason 1989 for bootstrap CLTs and Arcone and Gine 1990; Bose and Chatterjee 2003 for a more general treatment of bootstrapping M-estimators).

Step 1. In this step we establish that $\mathbb{W}_n^{(b)} \rightsquigarrow \mathbb{W}_\infty$ in probablity. Recall that

$$\mathbb{W}_n^{(b)} \triangleq (\Sigma_{1,n}^{(b)})^{-1}\mathbb{G}_n^{(b)} B_1^\intercal \left[Y_1 + H_{2,0}^\intercal \hat{\beta}_{2,0} + \left[H_{2,1}^\intercal \hat{\beta}_{2,1}\right]_+ - B\hat{\beta}_1\right] \\ + (\Sigma_{1,n}^{(b)})^{-1}\hat{\mathbb{P}}_n^{(b)} B_1^\intercal H_{2,0}^\intercal \sqrt{n}(\hat{\beta}_{2,0}^{(b)} - \hat{\beta}_{2,0}). \quad (3.34)$$

The second term in the above expression converges weakly to $\Sigma_{1,\infty}^{-1} P B_1^\intercal H_{2,0}^\intercal \mathbb{V}_{\infty,0}$. To see this notice $\hat{\mathbb{P}}_n^{(b)} B_1^\intercal H_{2,0}^\intercal \rightarrow P B_1^\intercal H_{2,1}^\intercal$ almost surely by the bootstrap LLN (Athreya 1983; Csorgo and Rosalsky 2003). Then use the bootstrap continuous mapping theorem (Kosorok 2008) and Slutsky's lemma to obtain the result. Turning attention to the first term on the right hand side of (3.34) we first note that the class of functions

$$\mathcal{F} \triangleq \left\{ B_1^\intercal \left[Y_1 + H_{2,0}^\intercal \beta_{2,0} + \left[H_{2,1}^\intercal \beta_{2,1}\right]_+ - B_1\beta_1\right] \,:\, \beta_2 \in \mathcal{B}(M, \beta_2^*),\, \beta_1 \in \mathcal{B}(M, \beta_1^*) \right\}$$

is a P-measurable BUEI class with square integrable envelope. This fact follows from the same arguments to the ones given above in the proof of Theorem 2.1, part 1. Also from part 1 of Theorem 2.1, it follows that $\hat{\beta}_1 \rightarrow \beta_1^*$ in probability and using standard results from least squares it follows that $\hat{\beta}_2 \rightarrow \beta_2^*$ in probability. Thus,

$$\mathbb{G}_n^{(b)} B_1^\intercal \left[Y_1 + H_{2,0}^\intercal \hat{\beta}_{2,0} + \left[H_{2,1}^\intercal \hat{\beta}_{2,1}\right]_+ - B_1\hat{\beta}_1\right] = \mathbb{G}_n^{(b)} B_1^\intercal \left[Y_1 + H_{2,0}^\intercal \beta_{2,0}^* + \left[H_{2,1}^\intercal \beta_{2,1}^*\right] - B_1\beta_1^*\right] + r_n$$

where

$$r_n = \mathbb{G}_n B_1^\intercal \left[H_{2,0}^\intercal(\hat{\beta}_{2,0} - \beta_{2,0}^*) + \left[H_{2,1}^\intercal \hat{\beta}_{2,1} \right]_+ - \left[H_{2,1}^\intercal \beta_{2,1}^* \right]_+ - B_1^\intercal(\hat{\beta}_1 - \beta_1^*) \right]$$

satisfies $P_M(||r_n|| > \epsilon) = o_P(1)$ for any $\epsilon > 0$ using the asymptotic equicontinuity of $\mathbb{G}_n^{(b)}$ in $l^\infty(\mathcal{F})$ (Kosorok 2008). Finally, applying the bootstrap CLT and Slutskly's lemma we have shown that $\mathbb{W}_n^{(b)} \rightsquigarrow \mathbb{W}_\infty$ in probability.

Step 2. In the second step we derive the limiting distribution of $(\Sigma_{1,n}^{(b)})^{-1}\hat{\mathbb{P}}_n^{(b)} B_1^\intercal \mathbb{U}_n^{(b)} 1_{T_n^{(b)}(H_{2,1})>\lambda_n}$. Using the definition of $\mathbb{U}_n^{(b)}$, Corollary (3.6.17), and Lemma (3.6.1), it follows that we can write the preceding quantity as

$$(\Sigma_{1,n}^{(b)})^{-1}\hat{\mathbb{P}}_n^{(b)} B_1^\intercal \mathbb{U}_n^{(b)} 1_{H_{2,1}^\intercal \beta_{2,1}^* > 0} + o_{P_\mathcal{M}}(1).$$

Furthermore, we can writhe the first term in the above display as

$$\begin{aligned}(\Sigma_{1,n}^{(b)})^{-1}\hat{\mathbb{P}}_n^{(b)} B_1^\intercal H_{2,1}^\intercal \sqrt{n}(\hat{\beta}_{2,1}^{(b)} - \hat{\beta}_{2,1}) 1_{(H_{2,1}^\intercal \hat{\beta}_{2,}^{(b)})(H_{2,1}^\intercal \hat{\beta}_{2,1})>0} 1_{H_{2,1}^\intercal \beta_{2,1}^* > 0} \\ + (\Sigma_{1,n}^{(b)})^{-1}\hat{\mathbb{P}}_n^{(b)} B_1^\intercal \mathbb{U}_n^{(b)} 1_{(H_{2,1}^\intercal \hat{\beta}_{2,}^{(b)})(H_{2,1}^\intercal \hat{\beta}_{2,1})<0} 1_{H_{2,1}^\intercal \beta_{2,1}^* > 0}. \quad (3.35)\end{aligned}$$

The event $\left\{ (H_{2,1}^\intercal \hat{\beta}_{2,}^{(b)})(H_{2,1}^\intercal \hat{\beta}_{2,1}) < 0 \right\}$ is contained in the event

$$\left\{ (H_{2,1}^\intercal \hat{\beta}_{2,}^{(b)})(H_{2,1}^\intercal \beta_{2,1}^*) < 0 \right\} \bigcup \left\{ (H_{2,1}^\intercal \hat{\beta}_{2,})(H_{2,1}^\intercal \beta_{2,1}^*) < 0 \right\}.$$

This containment, coupled with application of Corollary (3.6.17) show that the second term of (3.35) is $o_{P_\mathcal{M}}(1)$, while the first term is equal to

$$(\Sigma_{1,n}^{(b)})^{-1}\hat{\mathbb{P}}_n^{(b)} B_1^\intercal H_{2,1}^\intercal \sqrt{n}(\hat{\beta}_{2,1}^{(b)} - \hat{\beta}_{2,1}) 1_{H_{2,1}^\intercal \beta_{2,1}^* > 0} + o_{P_\mathcal{M}}(1).$$

The bootstrap continuous mapping theorem coupled with Slutsky's theorem shows that

$$(\Sigma_{1,n}^{(b)})^{-1}\hat{\mathbb{P}}_n^{(b)}B_1^{\intercal}H_{2,1}^{\intercal}\sqrt{n}(\hat{\beta}_{2,1}^{(b)}-\hat{\beta}_{2,1})1_{H_{2,1}^{\intercal}\beta_{2,1}^{*}>0}$$

converges to in distribution to $\Sigma_{1,\infty}^{-1}PB_1^{\intercal}H_{2,1}^{\intercal}\mathbb{V}_{\infty}1_{H_{2,1}^{\intercal}\beta_{2,1}^{*}>0}$ in probability.

Step 3. In this step we derive the limiting distribution of $(\Sigma_{1,n}^{(b)})^{-1}\hat{\mathbb{P}}_n^{(b)}B_1^{\intercal}\left[H_{2,1}^{\intercal}\sqrt{n}(\hat{\beta}_{2,1}^{(b)}-\hat{\beta}_{2,1})\right]_+$. Fix a vector $c\in\mathbb{R}^{\dim(B_1)}$, and let $M>0$ be arbitrary, arguments used in the course of proving Lemma (3.6.4) show that the class of functions

$$\mathcal{F}\triangleq\left\{c^{\intercal}B_1^{\intercal}\left[H_{2,1}^{\intercal}\eta\right]_+ : \mathcal{B}(M,0)\right\}$$

is a P-measurable, VC class with square integrable envelope and hence is strong Glivenko-Cantelli. Next write

$$\hat{\mathbb{P}}_n^{(b)}B_1^{\intercal}\left[H_{2,1}^{\intercal}\sqrt{n}(\hat{\beta}_{2,1}^{(b)}-\hat{\beta}_{2,1})\right]_+=PB_1^{\intercal}\left[H_{2,1}^{\intercal}\sqrt{n}(\hat{\beta}_{2,1}^{(b)}-\hat{\beta}_{2,1})\right]_++(\hat{\mathbb{P}}_n^{(b)}-P)B_1^{\intercal}\left[H_{2,1}^{\intercal}\sqrt{n}(\hat{\beta}_{2,1}^{(b)}-\hat{\beta}_{2,1})\right]_+.$$

The leading term converges weakly to $PB_1^{\intercal}\left[H_{2,1}^{\intercal}\mathbb{V}_{\infty}\right]_+$ in probability. Furthermore, for arbitrary $\epsilon>0$ the latter term in the above display satisfies

$$\begin{aligned}P_{\mathcal{M}}\left(||(\hat{\mathbb{P}}_n^{(b)}-P)B_1^{\intercal}\left[H_{2,1}^{\intercal}\sqrt{n}(\hat{\beta}_{2,1}^{(b)}-\hat{\beta}_{2,1})\right]_+||>\epsilon\right)&\leq P_{\mathcal{M}}\left(||\hat{\mathbb{P}}_n^{(b)}-P||_{\mathcal{F}}^{*}>\epsilon\right)\\&+P_{\mathcal{M}}\left(||\sqrt{n}(\hat{\beta}_{2,1}^{(b)}-\hat{\beta}_{2,1})||>M\right)\end{aligned}$$

the right hand of which can be arbitrarily small in probability for sufficiently large n and M. Thus, by Slutsky's theorem, it follows that $(\Sigma_{1,n}^{(b)})^{-1}\hat{\mathbb{P}}_n^{(b)}B_1^{\intercal}\left[H_{2,1}^{\intercal}\sqrt{n}(\hat{\beta}_{2,1}^{(b)}-\hat{\beta}_{2,1})\right]_+$ converges weakly to $\Sigma_{1,\infty}^{(b)}PB_1^{\intercal}\left[H_{2,1}^{\intercal}\mathbb{V}_{\infty}\right]_+$ in probability.

Step 4. In the step, we find the limiting distribution of the random process

$$\left(\Sigma_{1,n}^{(b)}\right)^{-1}\hat{\mathbb{P}}_n^{(b)} B_1^\intercal \left(\left[H_{2,1}^\intercal(\mathbb{V}_n^{(b)}+\gamma)\right]_+ - \left[H_{2,1}^\intercal\gamma\right]_+ - \left[H_{2,1}^\intercal \mathbb{V}_n^{(b)}\right]_+\right) 1_{T_n^{(b)}(H_{2,1})\le\lambda_n}. \quad (3.36)$$

First, notice that using Lemma (3.6.1) and the triangle inequality it follows that

$$\left|\left|\hat{\mathbb{P}}_n^{(b)} B_1^\intercal \left(\left[H_{2,1}^\intercal(\mathbb{V}_n^{(b)}+\gamma)\right]_+ - \Big[H_{2,1}^\intercal\gamma]_+ - \left[H_{2,1}^\intercal \mathbb{V}_n^{(b)}\right]_+\right) 1_{T_n^{(b)}(H_{2,1})\le\lambda_n}\right|\right| \le 2\hat{\mathbb{P}}_n^{(b)}||B_1||\,||H_{2,1}||\,||\mathbb{V}_n^{(b)}||$$

the right hand side of which is $O_{P_M}(1)$. Thus, by the bootstrap LLN and Slutksy's theorem we can write (3.36) as

$$\Sigma_{1,\infty}^{-1}\hat{\mathbb{P}}_n^{(b)} B_1^\intercal \left(\left[H_{2,1}^\intercal(\mathbb{V}_n^{(b)}+\gamma)\right]_+ - \left[H_{2,1}^\intercal\gamma\right]_+ - \left[H_{2,1}^\intercal \mathbb{V}_n^{(b)}\right]_+\right) 1_{T_n^{(b)}(H_{2,1})\le\lambda_n} + o_{P_\mathcal{M}}(1). \quad (3.37)$$

Let $\epsilon > 0$ be arbitrary and let B_ϵ be as defined in Lemma (3.6.16). Observe that

$$\left|\left|\hat{\mathbb{P}}_n^{(b)} B_1^\intercal \left(\left[H_{2,1}^\intercal(\mathbb{V}_n^{(b)}+\gamma)\right]_+ - \left[H_{2,1}^\intercal\gamma\right]_+ - \left[H_{2,1}^\intercal \mathbb{V}_n^{(b)}\right]_+\right) 1_{T_n^{(b)}(H_{2,1})\le\lambda_n} 1_{H_{2,1}^\intercal\beta_{2,1}^*\neq 0}\right|\right|$$

is bounded above by

$$2\hat{\mathbb{P}}_n^{(b)}||B_1||\,||H_{2,1}||\,||H_{2,1}||\,||\mathbb{V}_n^{(b)}||1_{H_{2,1}\notin B_\epsilon} + 2\hat{\mathbb{P}}_n^{(b)}||B_1||\,||H_{2,1}||\,||\mathbb{V}_n^{(b)}|| \sup_{h_{2,1}\in B_\epsilon} 1_{T_n(h_{2,1})\le\lambda_n} 1_{H_{2,1}^\intercal\beta_{2,1}^*\neq 0}$$

which is $o_{P_\mathcal{M}}(1)$ by Lemma (3.6.16). Thus, we can replace the indicator $1_{T_n^{(b)}(H_{2,1})\le\lambda_n}$ in (3.36) with $1_{H_{2,1}^\intercal\beta_{2,1}^*=0}$ up to a term of $o_{P_\mathcal{M}}(1)$. The class of functions given by

$$\mathcal{F} \triangleq \left\{B_1^\intercal \left(\left[H_{2,1}^\intercal(\delta+\gamma)\right]_+ - \left[H_{2,1}^\intercal\gamma\right]_+ - \left[H_{2,1}^\intercal\delta\right]_+\right) 1_{H_{2,1}^\intercal\beta_{2,1}^*=0} \,:\, \delta\in\mathcal{B}(M,0), \gamma\in\mathbb{R}^{\dim(\beta_{2,1}^*)}\right\}$$

is P-measurable BUEI class with square integrable envelope $2||B_1||\,||H_{2,1}||\,M$ and hence is strong Glivenko-Cantelli (see the arguments in the course of proving Lemma (3.6.4)). Thus, using bootstrap Glivenko-Cantelli results (Kosorok 2008) we have

$$\begin{aligned}
&||(\hat{\mathbb{P}}_n^{(b)} - P)B_1^{\intercal}\left(\left[H_{2,1}^{\intercal}(\mathbb{V}_n^{(b)} + \gamma)\right]_+ - \left[H_{2,1}^{\intercal}\gamma\right]_+ - \left[H_{2,1}^{\intercal}\mathbb{V}_n^{(b)}\right]_+\right)|| \\
&\qquad \le ||(\hat{\mathbb{P}}_n^{(b)} - P)||_{\mathcal{F}} + 2(\hat{\mathbb{P}}_n^{(b)} + P)||B_1||\,||H_{2,1}||\,||\mathbb{V}_n^{(b)}||1_{||\mathbb{V}_n^{(b)}||>M}
\end{aligned}$$

the right hand side of which can be made arbitrarily $P_{\mathcal{M}}$ small with arbitrarily high P probability for sufficiently large M and n. We have shown that (3.36) is equal to

$$\Sigma_{1,\infty}^{-1}PB_1^{\intercal}\left(\left[H_{2,1}^{\intercal}(\mathbb{V}_n^{(b)} + \gamma)\right]_+ - \left[H_{2,1}^{\intercal}\gamma\right]_+ - \left[H_{2,1}^{\intercal}\mathbb{V}_n^{(b)}\right]_+\right)1_{H_{2,1}^{\intercal}\beta_{2,1}^*=0} + o_{P_{\mathcal{M}}}(1).$$

The map from $\mathbb{R}^{\dim(\beta_{2,1}^*)}$ into $l^{\infty}(\mathbb{R}^{\dim(\beta_{2,1}^*)})$ given by

$$\delta \mapsto \Sigma_{1,\infty}^{-1}PB_1^{\intercal}\left(\left[H_{2,1}^{\intercal}(\delta + \gamma)\right]_+ - \left[H_{2,1}^{\intercal}\gamma\right]_+ - \left[H_{2,1}^{\intercal}\delta\right]_+\right)1_{H_{2,1}^{\intercal}\beta_{2,1}^*=0},$$

is uniformly continuous. Thus, by the preceding arguments and the bootstrap continuous mapping theorem (Kosorok 2008) it follows that (3.36) converges weakly to

$$\Sigma_{1,\infty}^{-1}PB_1^{\intercal}\left(\left[H_{2,1}^{\intercal}(\mathbb{V}_{\infty} + \gamma)\right]_+ - \left[H_{2,1}^{\intercal}\gamma\right]_+ - \left[H_{2,1}^{\intercal}\mathbb{V}_{\infty}\right]_+\right)1_{H_{2,1}^{\intercal}\beta_{2,1}^*=0}.$$

Step 5: The last step of the proof is to combine steps 1-4 to obtain the desired result by appeal to the Cramer-Wold device and the bootstrap continuous mapping theorem. □

Remark 3.6.20 (Plug-in Pretesting Approach)**.** A natural approach to constructing a confidence interval in a non-regular problem is "a plug-in pretesting approach." This approach, is similar in spirit to the ACI in that it partitions the training data using a series of hypothesis tests and uses different approximations on each partition. In particular, the plug-in pretesting estimator of $c^{\intercal}\sqrt{n}(\hat{\beta}_1 - \beta_1^*)$ is given by

$$c^{\intercal}\mathbb{W}_n + c^{\intercal}\Sigma_{1,n}^{-1}\mathbb{P}_n B_1^{\intercal}\mathbb{U}_n 1_{T_n(H_{2,1})>\lambda_n} + c^{\intercal}\Sigma_{1,n}^{-1}\mathbb{P}_n B_1^{\intercal}\left[H_{2,1}^{\intercal}\mathbb{V}_n\right]_+ 1_{T_n(H_{2,1})\leq\lambda_n}. \tag{3.38}$$

Confidence intervals are formed by bootstrapping this estimator. Under fixed alternatives, the plug-in pretesting estimator (PPE) is consistent. This consistency is established by recalling that $1_{T_n(h_{2,1})\leq\lambda_n} \to 1_{h_{2,1}^{\intercal}\beta_{2,1}^*=0}$ in probability and then comparing the last two terms of the PPE with (3.25).

However intuitive, the PPE does not perform well in small samples under some generative models (see the main body of the paper and the last section of this supplement). One explanation for this underperformance is that the PPE is not consistent under local alternatives. In particular, under a local generative model as described in (A4), it can be shown that the difference between the PPE and $c^{\intercal}\sqrt{n}(\hat{\beta}_1 - \delta_1^{(n)})$ is equal to

$$c^{\intercal}\Sigma_{1,n}^{-1}\mathbb{P}_n B_1^{\intercal}\left(\left[H_{2,1}^{\intercal}\left(\mathbb{V}_n^{(n)} + \gamma\right)\right]_+ - \left[H_{2,1}^{\intercal}\gamma\right]_+ - \left[H_{2,1}^{\intercal}\mathbb{V}_n^{(n)}\right]_+\right)1_{H_{2,1}^{\intercal}\beta_{2,1}^*=0} + o_{P_{n,\gamma}}(1), \tag{3.39}$$

which is does not vanish for any alternative γ for which $H_{2,1}^{\intercal}\gamma$ is not identically zero with probability one.

The expression in (3.39) offers yet another view of the ACI. In particular, one can view the last term of $\mathcal{U}(c)$ as approximating the supremum over local alternatives of the difference between the PPE and the target $c^{\intercal}\sqrt{n}(\hat{\beta}_1 - \delta_1^{(n)})$. In this way, the ACI can be thought of as a corrected version of the PPE where the correction is intended to safeguard against poor small sample performance.

3.7 Appendix II: Extension of the ACI to many stages and many treatments

In this appendix, we develop the ACI for the general case where there is an arbitrary finite number of stages of treatment, and an arbitrary finite number of treatment choices at each stage. We begin with a review of the Q-learning procedure in this setting.

3.7.1 Q-Learning in the general case

We observe an *i.i.d.* sample of trajectories $\{\mathcal{T}_i\}_{i=1}^{n}$ drawn from a fixed but unknown distribution P. Each trajectory is of the form

$$\mathcal{T} = (X_1, A_1, Y_1, X_2, A_2, Y_2, \ldots, X_T, A_T, Y_T), \tag{3.40}$$

being comprised of patient measurement X_t, assigned treatment A_t, and observed response Y_t for $t = 1, 2, \ldots, T$. For each decision point t the assigned treatment A_t takes values in the set $\{1, 2, \ldots, K_t\}$. As in the two-stage setting, we let H_t denote a concise summary of patient history at time t. More precisely, $H_1 \triangleq \Psi_1(X_1)$ and $H_t \triangleq \Psi_t(X_1, A_1, Y_1, \ldots, X_{t-1}, A_{t-1}, Y_{t-1}, X_t)$ for $t = 2, 3, \ldots, T$ for known functions, Ψ_t. The form of the working model for the Q-function is of the same form as in Section 3 of the main body of the paper. For each t we use the model

$$Q_t(h_t, a_t; \beta_t) \triangleq \beta_{t,0}^{\intercal} h_{t,0} + \sum_{i=1}^{K_t} \beta_{t,i}^{\intercal} h_{t,1} 1_{a_t = i}, \tag{3.41}$$

where $\beta_t \triangleq (\beta_{t,0}^{\intercal}, \beta_{t,1}^{\intercal}, \beta_{t,2}^{\intercal}, \ldots, \beta_{t,K_t}^{\intercal})^{\intercal}$. For the purpose of identifiability, we assume that the vector of coefficients β_t satisfies a zero-sum constraint. That is, for each

$j = 1, 2, \ldots, dim(\beta_{t,1})$ the vector β_t satisfies $\sum_{i=1}^{K_t} \beta_{t,i,j} = 0$. As in the two stage setting, the form of the working model implies that when $h_{t,1}^{\intercal}\beta_{t,i} - \max_{j \neq i} h_{t,1}^{\intercal}\beta_{t,j} \approx 0$ for some $1 \leq i \leq K_t$, then at least two treatments are approximately optimal for a patient with history $H_{t,1} = h_{t,1}$. That is, there is *not* a unique best treatment for a patient with history $H_{t,1} = h_{t,1}$. On the other hand, if $\left| h_{t,1}^{\intercal}\beta_{t,i} - \max_{j \neq i} h_{t,1}^{\intercal}\beta_{t,j} \right| \gg 0$ for all $1 \leq i \leq K_t$, then the working model implies that *exactly* one treatment is best for a patient with history $H_{t,1} = h_{t,1}$. Once a working model has been specified, the Q-learning algorithm can be applied to estimate the optimal DTR. The Q-learning algorithm is a follows:

1. Regress Y_T on H_T and A_T using the working model (3.41) to obtain:

$$\hat{\beta}_T \triangleq \arg\min_{\beta_T} \mathbb{P}_n \left(Y_T - Q_T(H_T, A_T; \beta_T) \right)^2 \tag{3.42}$$

   and subsequently the approximation $Q_T(h_T, a_T; \hat{\beta}_T)$ to the conditional mean $Q_T(h_T, a_T)$.

2. (a) Recursively, define the predicted future reward following the optimal policy as:

$$\begin{aligned} \tilde{Y}_t &\triangleq Y_t + \max_{a_{t+1} \in \{1,2,\ldots,K_{t+1}\}} Q_{t+1}\left(H_{t+1}, a_{t+1}; \hat{\beta}_{t+1}\right) && (3.43) \\ &= Y_t + H_{t+1,0}^{\intercal}\hat{\beta}_{t+1,0} + \max_{1 \leq i \leq K_{t+1}} H_{t+1,1}^{\intercal}\hat{\beta}_{t+1,i} && (3.44) \end{aligned}$$

   for $t = T-1, T-2, \ldots, 1$.

   (b) Regress $\tilde{Y}_t$ on H_t and A_t using the working model (3.41) to obtain $\hat{\beta}_t \triangleq \arg\min_{\beta_t} \mathbb{P}_n \left(\tilde{Y}_t - Q_t(H_t, A_t; \beta_t) \right)^2$.

3. Define the estimated optimal DTR $\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_T)$ so that

$$\hat{\pi}_t(h_t) \triangleq \arg \max_{a_t \in \{1,2,\ldots,K_t\}} Q_t(h_t, a_t; \hat{\beta}_t). \tag{3.45}$$

When $T = 2$ the above procedure is equivalent to the two stage Q-learning algorithm given in Section 3 of the main body of the paper.

Our aim is to use the ACI to construct a confidence interval for $c^\intercal \beta_1^*$ where c is an arbitrary vector of constants. The definition of β_1^* is given inductively. Define

$$\beta_T^* \triangleq \arg\min_{\beta_T} P\left(Y_T - Q_T(H_T, A_T; \beta_T)\right)^2. \tag{3.46}$$

For $t = T-1, T-2, \ldots, 1$ define

$$\begin{aligned} \tilde{Y}_t^* &\triangleq Y_t + H_{t+1,0}^\intercal \beta_{t+1,0}^* + \max_{1 \le i \le K_{t+1}} H_{t+1,1}^\intercal \beta_{t+1,i}^*, & (3.47) \\ \beta_t^* &\triangleq \arg\min_{\beta_t} P\left(\tilde{Y}_T^* - Q_t(H_t, A_t; \beta_t)\right)^2. & (3.48) \end{aligned}$$

We focus on the problem of constructing a confidence interval for a linear combination of the first stage coefficients β_1^* since building a confidence interval for, say $c^\intercal \beta_t^*$, is equivalent to building a confidence interval for the first stage of a $T-t+1$ stage trial. That is, one can always view the tth stage as the first stage of a shorter $T-t+1$ stage trial. Information collected prior to the tth stage can be treated as baseline (pre-randomization) information in this shorter trial.

3.7.2 ACI in the general case

The ACI in the general case is conceptually the same as the two stage case. Non-regularity in $\sqrt{n}(\hat{\beta}_t - \beta_t^*)$ arises whenever there are two or more equally best treatments at *any* future stage of treatment $s > t$ for a non-null subset of patient histories. The ACI works by constructing smooth upper and lower bounds on $\sqrt{n}(\hat{\beta} - \beta_t^*)$ and then bootstrapping these bounds to construct confidence intervals. As in the two stage case, these bounds are asymptotically equivalent to taking the supremum (infimum) over all local alternatives to the true generative distribution.

In order to develop the ACI in this general setting, we generalize the notation given in the main body of the paper. Define $B_t \triangleq (H_{t,0}^{\intercal}, H_{t,1}^{\intercal} 1_{A_t=1}, \ldots, H_{t,1}^{\intercal} 1_{A_t=K_t})^{\intercal}$ so that instances of B_t form the rows of the design matrix used in the tth stage regression. Further, define $\Sigma_{t,n} \triangleq \mathbb{P}_n B_t^{\intercal} B_t$. The limiting distribution of $\sqrt{n}(\hat{\beta}_t - \beta_t^*)$ depends abruptly on the frequency of patients for which there are multiple equally optimal best treatments at a future stage. Consequently, the set

$$\mathcal{A}_t^*(h_{t,1}) \triangleq \left\{ \arg \max_{1 \le i \le K_t} h_{t,1}^T \beta_{t,i}^* \right\} \tag{3.49}$$

of equally optimal treatments at stage t for a patient with history $H_{t,1} = h_{t,1}$, is relevant for the development of asymptotic theory. Notice that $\mathcal{A}_t^*(h_{t,1})$ is a singleton when there is exactly one best treatment for a patient with history $h_{t,1}$. As in the two stage case, we will need to estimate $\mathcal{A}_t^*(h_{t,1})$. The estimator we use is based on the following test statistics:

$$T_{t,n,i}(h_{t,1}) \triangleq \frac{n \left(h_{t,1}^{\intercal} \hat{\beta}_{t,i} - \max_{j \neq i} h_{t,1}^{\intercal} \hat{\beta}_{t,j} \right)^2}{h_{t,1}^{\intercal} \hat{\zeta}_{t,i} h_{t,1}}$$

where $\hat{\zeta}_i$ is the usual plug-in estimator of $n\text{Cov}(\hat{\beta}_{t,i} - \hat{\beta}_{t,k})$ evaluated at $k = \hat{k}_i$ where $\hat{k}_i = \arg\max_{j \neq i} h_{t,1}^{\intercal} \hat{\beta}_j$ (we are acting as if the maximal index is fixed a priori).
**** the reader will not know how we define $\hat{\zeta}_i$ except when $t = T$. –how did we define in simulation? We should use this definition. Did we use $\hat{\zeta}_{t,i}$ equal to a *difference* of

submatrices from $\Sigma_{t,n}^{-1}$ or did we use $\hat{\zeta}_{t,i}$ equal to the submatrix corresponding to i? ****

The statistic, $\min_i T_{t,n,i}(h_{t,1})$, should be large when there is exactly one best treatment for a patient with history $H_{t,1} = h_{t,1}$. On the other hand, $\min_i T_{t,n,i}(h_{t,1})$ should be small if treatment i is an optimal treatment for a patient with history $h_{t,1}$ *and* there is more than one best treatment. Thus, a natural estimator of $\mathcal{A}_t^*(h_{t,1})$ is

$$\hat{\mathcal{A}}_t(h_{t,1}) = \begin{cases} \{i \,:\, T_{t,n,i}(h_{t,1}) \leq \lambda_n\} & \text{if } \min_i T_{t,n,i}(h_{t,1}) \leq \lambda_n \\ \arg\max_{1\leq i\leq K_t} h_{t,1}^{\intercal}\hat{\beta}_{t,i} & \text{if } \min_i T_{t,n,i}(h_{t,1}) > \lambda_n. \end{cases}$$

The merits and genesis of this statistic were discussed in the main body of the paper. Under the regularity conditions given in the next section, it follows that $\hat{\mathcal{A}}_t(h_{t,1})$ is a consistent estimator of $\mathcal{A}_t^*(h_{t,1})$.

It is also useful to define generalizations of the $\mathbb{V}_{t,n}$'s. Let N_t be the dimension of $(\beta_{t,1}^{*\intercal}, \beta_{t,2}^{*\intercal}, \ldots, \beta_{t,K_t}^{*\intercal})^{\intercal}$, that is $N_t = K_t \dim(\beta_{t,1}^*)$ (note this does not include the dimension of $\beta_{t,0}^*$). Define $\mathbb{V}_{T,n} \triangleq \sqrt{n}(\hat{\beta}_T - \beta_T^*)$ and $\mathbb{V}_{T,n,i} \triangleq \sqrt{n}(\hat{\beta}_{T,i} - \beta_{T,i}^*)$ and for any $\gamma_T \in \mathbb{R}^{N_T}$ define

$$\begin{aligned} \mathbb{V}_{T-1,n}(\gamma_T) \triangleq \mathbb{W}'_{T-1,n} &+ \Sigma_{T-1,n}^{-1}\mathbb{P}_n B_{T-1}^{\intercal} H_{T,0}^{\intercal}\mathbb{V}_{T,n,0} + \Sigma_{T-1,n}^{-1}\mathbb{P}_n B_{T-1}^{\intercal}\mathbb{U}_{T,n}1_{\#\hat{\mathcal{A}}_T(H_{T,1})=1} \\ &+ \Sigma_{T-1,n}^{-1}\mathbb{P}_n B_{T-1}^{\intercal} \max_{i\in\hat{\mathcal{A}}_T(H_{T,1})} H_{T,1}^{\intercal}\mathbb{V}_{T,n,i}1_{\#\hat{\mathcal{A}}_T(H_{T,1})>1} \\ + \Sigma_{T-1,n}^{-1}\mathbb{P}_n B_{T-1}^{\intercal} &\left[\max_{1\leq i\leq K_T} H_{T,1}^{\intercal}(\mathbb{V}_{T,n,i} - \gamma_{T,i}) - \max_{1\leq i\leq K_T} H_{T,1}^{\intercal}\gamma_{T,i} - \max_{i\in\hat{\mathcal{A}}_T(H_{T,1})} H_{T,1}^{\intercal}\mathbb{V}_{T,n,i}\right] \\ &\times 1_{\#\hat{\mathcal{A}}_T(H_{T,1})>1}1_{\gamma_T\in\mathcal{C}_{T,n}(H_{T,1})}, \end{aligned}$$

where

$$\begin{aligned}
\mathbb{W}'_{t,n} &= \Sigma_{t,n}^{-1}\sqrt{n}\mathbb{P}_n B_t^{\intercal}\left\{Y_t + H_{t+1,0}^{\intercal}\beta_{t+1,0}^{*} + \max_{1\le i\le K_{t+1}} H_{t+1,1}^{\intercal}\beta_{t+1,i}^{*} - B_t\beta_t^{*}\right\} \\
\mathbb{U}_{t,n} &= \sqrt{n}\left(\max_{1\le i\le K_t} H_{t,1}^{\intercal}\hat{\beta}_{t,i} - \max_{1\le i\le K_t} H_{t,1}^{\intercal}\beta_{t,i}^{*}\right) \\
\mathcal{C}_{t,n}(h_{t,1}) &\triangleq \left\{\gamma_t \in \mathbb{R}^{N_t}, i \in \hat{\mathcal{A}}_t(h_{t,1}) \,:\, \frac{\left(h_{t,1}^{\intercal}(\mathbb{V}_{t,n,i} - \gamma_{t,i}) - \max_{j\neq i} h_{t,1}^{\intercal}(\mathbb{V}_{t,n,j} - \gamma_{t,j})\right)^2}{h_{t,1}^{\intercal}\hat{\zeta}_{t,i}h_{t,1}} \le \lambda_n\right\}.
\end{aligned}$$

Notice that $\mathbb{V}_{T-1,n}(\sqrt{n}\beta_{T-1}^{*}) = \mathbb{V}_{T-1,n}$. If $T = 2$, then $\mathcal{U}_1(c) = \sup_{\gamma_T} c^{\intercal}\mathbb{V}_{t,n}(\gamma_T)$. More generally, for any $t < T-1$ and $\Gamma_{t+1} = (\gamma_{t+1}^{\intercal}, \gamma_{t+2}^{\intercal}, \ldots, \gamma_T^{\intercal})^{\intercal} \in \mathbb{R}^{\sum_{k=t+1}^{T} N_k}$ define

$$\begin{aligned}
\mathbb{V}_{t,n}(\Gamma_{t+1}) \triangleq \mathbb{W}'_{t,n} &+ \Sigma_{t,n}^{-1}\mathbb{P}_n B_t^{\intercal} H_{t+1,0}^{\intercal}\mathbb{V}_{t+1,n,0}(\Gamma_{t+2}) \\
&+ \Sigma_{t,n}^{-1}\mathbb{P}_n B_t^{\intercal} \max_{i\in\hat{\mathcal{A}}_{t+1}(H_{t+1,1})} H_{t+1,1}^{\intercal}\mathbb{V}_{t+1,n,i}(\Gamma_{t+2}) \\
&+ \Sigma_{t,n}^{-1}\mathbb{P}_n B_t^{\intercal}\left(\mathbb{U}_{t+1,n} - \max_{i\in\hat{\mathcal{A}}_{t+1}(H_{t+1,1})} H_{t+1,1}^{\intercal}\mathbb{V}_{t+1,n,i}\right) 1_{\#\hat{\mathcal{A}}_{t+1}(H_{t+1,1})=1} \\
&\Sigma_{t,n}^{-1}\mathbb{P}_n B_t^{\intercal}\Bigg(\max_{1\le i\le K_{t+1}} H_{t+1,1}^{\intercal}\left(\mathbb{V}_{t+1,n,i}(\Gamma_{t+2}) + \gamma_{t+1}\right) - \max_{1\le i\le K_{t+1}} H_{t+1,1}^{\intercal}\gamma_{t+1,i} \\
&- \max_{i\in\hat{\mathcal{A}}_{t+1}(H_{t+1,1})} H_{t+1,1}^{\intercal}\mathbb{V}_{t+1,n,i}(\Gamma_{t+2})\Bigg) 1_{\#\hat{\mathcal{A}}_{t+1}(H_{t+1,1})>1} 1_{\gamma_{t+1}\in\mathcal{C}_{t+1}(H_{t+1,1})}.
\end{aligned}$$

The upper bound on $c^{\intercal}\sqrt{n}(\hat{\beta}_t - \beta_t^{*})$ used to construct a confidence interval for $c^{\intercal}\beta_t^{*}$ is given by $\mathcal{U}_t(c) \triangleq \sup_{\Gamma_{t+1}\in\mathbb{R}^{\sum_{k=t+1}^{T} N_k}} c^{\intercal}\mathbb{V}_{t,n}(\Gamma_{t+1})$. Similarly, the lower bound is obtained by replacing the sup with an inf.

3.7.2.1 Example: ACI for three stages

To illustrate the ACI for the general case and solidify the ideas presented in the preceding section, we provide the bounds for the case where there are three stages of treatment and an arbitrary number of treatments at each stage. Thus, $T = 3$ and $\mathbb{V}_{3,n} \triangleq \sqrt{n}(\hat{\beta}_3 - \beta_3^{*})$. Since $\mathbb{V}_{3,n}$ is the usual least squares estimator, it follows under

(A1)-(A2) (see below) that $\mathbb{V}_{3,n}$ is regular and its limiting distribution is normal. The process $\mathbb{V}_{2,n}(\gamma_3)$ is indexed by $\gamma_3 \in \mathbb{R}^{N_3}$ and is defined as follows

$$\begin{aligned}
\mathbb{V}_{2,n}(\gamma_3) \triangleq \mathbb{W}'_{2,n} + \Sigma_{2,n}^{-1}\mathbb{P}_n B_2^\intercal H_{3,0}^\intercal \mathbb{V}_{3,n,0} + \Sigma_{2,n}^{-1}\mathbb{P}_n B_2^\intercal \mathbb{U}_{3,n} 1_{\#\hat{\mathcal{A}}_3(H_{3,1})=1} \\
+ \Sigma_{2,n}^{-1}\mathbb{P}_n B_2^\intercal \max_{i\in\hat{\mathcal{A}}_3(H_{3,1})} H_{3,1}^\intercal \mathbb{V}_{3,n,i} 1_{\#\hat{\mathcal{A}}_3(H_{3,1})>1} \\
+ \Sigma_{2,n}^{-1}\mathbb{P}_n B_2^\intercal \left[\max_{1\le i\le K_3} H_{3,1}^\intercal \left(\mathbb{V}_{3,n,i} + \gamma_{3,i}\right) - \max_{1\le i\le K_3} H_{3,1}^\intercal \gamma_{3,i} - \max_{i\in\hat{\mathcal{A}}_3(H_{3,1})} H_{3,1}^\intercal \mathbb{V}_{3,n,i} \right] \\
1_{\#\hat{\mathcal{A}}_3(H_{3,1})>1} 1_{\gamma_3\in\mathcal{C}_{3,n}(H_{3,1})}.
\end{aligned}$$

A confidence bound for $c^\intercal\beta_2^*$ is formed using the boodstrap distribution of bounds $\mathcal{U}_2(c) \triangleq \sup_{\gamma_3\in\mathbb{R}^{N_3}} c^\intercal \mathbb{V}_{2,n}(\gamma_3)$ and $\mathcal{L}_2(c) \triangleq \inf_{\gamma_3\in\mathbb{R}^{N_3}} c^\intercal \mathbb{V}_{2,n}(\gamma_3)$.

To form a confidence interval for the first stage coefficients, e.g. $c^\intercal\beta_1^*$ we use the process $\mathbb{V}_{1,n}((\gamma_2,\gamma_3))$ which is indexed by $\gamma_2,\gamma_3 \in \mathbb{R}^{N_2+N_3}$. The definition of $\mathbb{V}_{n,1}((\gamma_2,\gamma_3))$ is given by

$$\begin{aligned}
\mathbb{V}_{1,n}((\gamma_2,\gamma_3)) \triangleq \mathbb{W}'_{1,n} + \Sigma_{1,n}^{-1}\mathbb{P}_n B_1^\intercal H_{2,0}^\intercal \mathbb{V}_{2,n,0}(\gamma_3) + \Sigma_{1,n}^{-1}\mathbb{P}_n B_1^\intercal \max_{i\in\hat{\mathcal{A}}_2(H_{2,1})} H_{2,1}^\intercal \mathbb{V}_{2,n,i}(\gamma_3) \\
+ \Sigma_{1,n}^{-1}\mathbb{P}_n B_1^\intercal \left(\mathbb{U}_{2,n} - \max_{i\in\hat{\mathcal{A}}_2(H_{2,1})} H_{2,1}^\intercal \mathbb{V}_{2,n,i} \right) 1_{\#\hat{\mathcal{A}}_{2,1}(H_{2,1})=1} \\
+\Sigma_{1,n}^{-1}\mathbb{P}_n B_1^\intercal \left(\max_{1\le i\le K_2} H_{2,1}^\intercal \left(\mathbb{V}_{2,n,i}(\gamma_3) + \gamma_{2,i}\right) - \max_{1\le i\le K_2} H_{2,1}^\intercal \gamma_{2,i} - \max_{i\in\hat{\mathcal{A}}_2(H_{2,1})} H_{2,1}^\intercal \mathbb{V}_{2,n,i}(\gamma_3) \right) \\
1_{\#\hat{\mathcal{A}}_2(H_{2,1})>1} 1_{\gamma_2\in\mathcal{C}_{2,n}(H_{2,1})}.
\end{aligned}$$

Thus, the upper and lower bounds used for constructing a confidence interval for $c^\intercal\beta_1^*$ are given by $\mathcal{U}_1(c) \triangleq \sup_{\gamma_2,\gamma_3\in\mathbb{R}^{N_2+N_2}} c^\intercal \mathbb{V}_{1,n}((\gamma_2,\gamma_3))$ and $\mathcal{L}_1(c) \triangleq \inf_{\gamma_2,\gamma_3\in\mathbb{R}^{N_2+N_3}} c^\intercal \mathbb{V}_{1,n}((\gamma_2,\gamma_3))$. The form of $\mathbb{V}_{2,n}(\gamma_3)$ and $\mathbb{V}_{1,n}((\gamma_2,\gamma_3))$ show that computing the bounds $\mathcal{U}_1(c)$ and $\mathcal{L}_1(c)$ require optimizing piecewise linear objective functions. Since these piecewise linear objectives are non-convex (non-concave) the resultant optimization problem is, to the best of our knowledge, a mixed integer program. A simple stochastic approximation is given in Section 1.4 of this supplement.

3.7.3 Properties of the ACI in the general case

In this section, we state the general case analogs of the theorems given in the main body of the paper. In particular, these results state that the ACI provides asymptotically valid confidence intervals under mild regularity conditions. In addition, under further assumptions, it can be shown that the ACI delivers asymptotically exact coverage.

We will make the following moment assumptions.

(A1) The histories H_t, features B_t, and outcomes Y_t, satisfy the moment inequalities $P||H_t||^2\,||B_{t-1}||^2 < \infty$ for all $t = 2, 3, \ldots, T$, and $PY_t^2\,||B_t||^2 < \infty$ for all $t = 1, 2, \ldots, T$.

(A2) Define:

1. $\Sigma_{t,\infty} \triangleq PB_t^{\intercal}B_t$;
2. $g_T(B_T, Y_T; \beta_T^*) \triangleq B_T^{\intercal}(Y_T - B_T\beta_T^*)$;
3. $g_t(B_t, Y_t, H_{t+1}; \beta_t^*) \triangleq B_t^{\intercal}\left(Y_T + H_{t+1,0}^{\intercal}\beta_{t+1,0}^* + \max_{k\in\mathcal{A}_{t+1}^*} H_{t+1,1}^{\intercal}\beta_{t+1,k}^* - B_t\beta_t^*\right)$;

then the matrices $\Sigma_{t,\infty}$ for $t = 1, \ldots, T$, and $\Omega \triangleq Cov(g_1, g_2, \ldots, g_T)$ are strictly positive definite.

(A3) The sequence λ_n tends to infinity and satisfies $\lambda_n = o(n)$.

These assumptions are quite mild requiring the kind of moment and collinearity constraints which are often encountered in multiple regression. The last assumption (A3) concerns a user-controlled parameter and is thus readily satisfied. Let $\mathbb{V}_{t,\infty}(\Gamma_{t+1})$ denote the limiting process of $\mathbb{V}_{t,n}(\Gamma_{t+1})$ which is indexed by $\Gamma_{t+1} \in \mathbb{R}^{\prod_{k=t+1}^{T} N_k}$. We write $\mathbb{V}_{t,\infty}$ to denote the limiting distribution of $\mathbb{V}_{n,t}$.

Theorem 3.7.1 (Validity of Population Bounds). *Assume (A1)-(A3) and fix $c \in \mathbb{R}^{\dim(\beta_t)^*}$.*

1. *The limiting distribution $c^\intercal\sqrt{n}(\hat{\beta}_t - \beta_t^*)$ is given by:*

$$c^\intercal \mathbb{W}'_{t,\infty} + c^\intercal \Sigma_{t,\infty}^{-1} P B_t^\intercal H_{t+1,0}^\intercal \mathbb{V}_{t+1,\infty,0} + c^\intercal \Sigma_{t,\infty}^{-1} P B_t^\intercal \max_{i \in \mathcal{A}_{t+1}^*(H_{t+1,1})} H_{t+1}^\intercal \mathbb{V}_{t+1,\infty,i}.$$

2. *The limiting distribution $\mathcal{U}_{T-1}(c)$ is given by*

$$c^\intercal \mathbb{W}'_{T-1,\infty} + c^\intercal \Sigma_{T-1,\infty}^{-1} P B_{T-1}^\intercal H_{T,0}^\intercal \mathbb{V}_{T,\infty,0} + c^\intercal \Sigma_{T-1,\infty}^{-1} P B_{T-1}^\intercal \max_{i \in \mathcal{A}_T^*(H_{T,1})} H_{T,1}^\intercal \mathbb{V}_{T,\infty,i} 1_{\#\mathcal{A}_T^*(H_{T,1})=1}$$
$$+ \sup_{\gamma_T \in \mathbb{R}^{N_T}} c^\intercal \Sigma_{T-1,\infty}^{-1} P B_{T-1}^\intercal \left(\max_{i \in \mathcal{A}_T^*(H_{T,1})} H_{T,1}^\intercal (\mathbb{V}_{T,\infty,i} + \gamma_{T,i}) - \max_{i \in \mathcal{A}_T^*(H_{T,1})} H_{T,1}^\intercal \gamma_{T,i} \right) 1_{\#\mathcal{A}_T^*(H_{T,1})>1}.$$

3. *For $t < T-1$, the limiting distribution of $\mathcal{U}_t(c)$ is given (recursively) by:*

$$c^\intercal \mathbb{W}'_{t,\infty} + \sup_{\Gamma_{t+1}} \Bigg\{ c^\intercal \Sigma_{t,\infty}^{-1} P B_t^\intercal H_{T+1,0}^\intercal \mathbb{V}_{t+1,\infty,0}(\Gamma_{t+2})$$
$$+ c^\intercal \Sigma_{t,\infty}^{-1} P B_t^\intercal \max_{i \in \mathcal{A}_{t+1}^*(H_{t,1})} H_{t+1,1}^\intercal \mathbb{V}_{t+1,\infty,i}(\Gamma_{t+2}) 1_{\#\mathcal{A}_{t+1}^*(H_{t,1})=1}$$
$$+ c^\intercal \Sigma_{t,\infty}^{-1} P B_t^\intercal \left(\max_{i \in \mathcal{A}_T^*(H_{T,1})} H_{t+1,1}^\intercal \left(\mathbb{V}_{t+1,\infty,i}(\Gamma_{t+2}) + \gamma_{t+1}\right) - \max_{i \in \mathcal{A}_T^*(H_{T,1})} H_{t+1,1}^\intercal \gamma_{t+1,i} \right) 1_{\#\mathcal{A}_{t+1}^*(H_{t,1})>1} \Bigg\}$$

When $T = 2$, these limiting distributions are equal in law to the limiting distributions of $\mathcal{U}(c)$ and $\mathcal{L}(c)$ given in Section 2 of the main body of the paper. The preceding theorem shows that the limiting distribution of $\mathcal{U}_t(c)$ is stochastically larger than that of $c^\intercal\sqrt{n}(\hat{\beta}_t - \beta_t^*)$. A similar result can be stated in terms of $\mathcal{L}_t(c)$ by replacing the sup by an inf in the preceding theorem. The theorem is proved recursively using the results proved for the two stage case and then repeatedly invoking the continuous mapping theorem.

In order to form a confidence interval, the bootstrap distributions of $\mathcal{U}_t(c)$ and $\mathcal{L}_t(c)$, which we denote by $\mathcal{U}_t^{(b)}(c)$ and $\mathcal{L}_t^{(b)}(c)$, are used. The next result states that the bootstrap bounds are asymptotically consistent.

Theorem 3.7.2. *Assume (A1)-(A3) and fix* $c \in \mathbb{R}^{\dim(\beta_t^*)}$. *Then for* $(\mathcal{U}_t(c), \mathcal{L}_t(c))$ *and* $(\mathcal{U}_t^{(b)}(c), \mathcal{L}_t^{(b)}(c))$ *converge to the same limiting distribution in probability. That is,*

$$\sup_{\nu \in BL_1(\mathbb{R}^2)} \left| \mathbb{E}\nu\left((\mathcal{U}_t(c), \mathcal{L}_t(c))\right) - \mathbb{E}_M \nu\left(\left(\mathcal{U}_t^{(b)}(c), \mathcal{L}_t^{(b)}(c)\right)\right) \right|$$

converges to zero in probability.

Corollary 3.7.3. *Assume (A1)-(A3) and fix* $c \in \mathbb{R}^{\dim(\beta_1^*)}$. *Let* $\hat{u}$ *denote the* $1 - \alpha/2$ *quantile of* $\mathcal{U}_1^{(b)}(c)$ *and* $\hat{l}$ *denote the* $\alpha/2$ *quantile of* $\mathcal{L}^{(b)}(c)$. *Then*

$$P_M\left(c^{\intercal}\hat{\beta}_1 - \hat{u}/\sqrt{n} \le c^{\intercal}\beta_1^* \le c^{\intercal}\hat{\beta}_1 - \hat{l}/\sqrt{n}\right) \ge 1 - \alpha + o_P(1).$$

Furthermore, if $P\left(\min_i \left|H_{t,1}^{\intercal}\beta_{t,i}^* - \max_{j \ne i} H_{t,1}^{\intercal}\beta_{t,j}^*\right| = 0\right) = 0$ *for all* $t = 2, 3, \ldots, T$, *then the above inequality can be strengthened to equality.*

The preceding result shows that the ACI can be used to construct valid confidence intervals regardless of the underlying parameters or generative model. In addition, when there is almost always a unique best treatment, then the ACI delivers asymptotically exact confidence intervals.

3.8 Appendix III: Proofs for the ACI in more than two stages and more than two treatments

Lemma 3.8.1 (Toy Inequality). *Let $a, b \in \mathbb{R}^d$ be arbitrary. Then*

$$\max_{1\le i\le d}(a_i + b_i) - \max_{1\le i\le d} b_i \le \max_{1\le i\le d} a_i \le ||a||_\infty .$$

Proof. Let $i^* \in \arg\max_i(a_i + b_i)$, then

$$\max_{1\le i\le d}(a_i + b_i) - \max_{1\le i\le d} b_i \le a_{i^*} \le \max_{1\le i\le K_t} a_i \le \max_i |a_i| = ||a||_\infty .$$

□

Lemma 3.8.2. *Consistency of $\hat\beta_t$. Assume (A1)-(A2), then for each t it follows that $\sqrt{n}(\hat\beta_t - \beta_t^*) = O_P(1)$, and $P_{\mathcal{M}}\left(\left|\left|\sqrt{n}(\hat\beta_t^{(b)} - \hat\beta_t)\right|\right| > L\right) = o_P(1)$ as L tends to infinity, where $P_{\mathcal{M}}$ denotes probability with respect to the bootstrap weights.*

Proof. The proof proceeds by backwards induction. The base case follows immediately since $\sqrt{n}(\hat\beta_T - \beta_T^*)$ is the usual least squares estimator and hence is asymptotically normal and thus $O_P(1)$. Suppose as the inductive step that $\sqrt{n}(\hat\beta_{t+1} - \beta_{t+1}^*) = O_P(1)$, the result follows if we can establish that $\sqrt{n}(\hat\beta_t - \beta_t^*) = O_P(1)$. Note that $\sqrt{n}(\hat\beta_t - \beta_t^*)$ can be decomposed as follows

$$\sqrt{n}(\hat\beta_t - \beta_t^*) = \mathbb{W}'_{t,n} + \Sigma_{t,n}^{-1}\mathbb{P}_n B_t^\intercal H_{t+1,0}^\intercal \sqrt{n}(\hat\beta_{t+1,0} - \beta_{t+1,0}^*) + \Sigma_{t,n}^{-1}\mathbb{P}_n B_t^\intercal \mathbb{U}_{t+1,n}. \quad (3.50)$$

The proof that the sum of the first two terms is $O_P(1)$ is immediate and omitted. Consider the third term.

$$\left|\left|\mathbb{P}_n B_t^\intercal \mathbb{U}_{t+1,n}\right|\right| \;=\; \left|\left|\sqrt{n}\mathbb{P}_n B_t^\intercal \left(\max_{1\le i\le K_{t+1}} H_{t+1,1}^\intercal \hat\beta_{t+1,i} - \max_{1\le i\le K_{t+1}} H_{t+1,1}^\intercal \beta_{t+1,i}^*\right)\right|\right|$$

$$
\begin{aligned}
&\leq \mathbb{P}_n ||B_t|| \sqrt{n} \left| \max_{1 \leq i \leq K_{t+1}} H_{t+1,1}^{\intercal} \hat{\beta}_{t+1,i} - \max_{1 \leq i \leq K_{t+1}} H_{t+1,1}^{\intercal} \beta_{t+1,i}^{*} \right| \\
&\leq \mathbb{P}_n ||B_t|| \max_{1 \leq i \leq K_{t+1}} \left| H_{t+1,1}^{\intercal} \sqrt{n} (\hat{\beta}_{t+1,i} - \beta_{t+1,i}^{*}) \right| \\
&\leq \mathbb{P}_n ||B_t|| \, ||H_{t+1,1}|| \max_{1 \leq i \leq K_{t+1}} \left| \left| \sqrt{n} (\hat{\beta}_{t+1,i} - \beta_{t+1,i}^{*}) \right| \right| \\
&= O_P(1),
\end{aligned}
$$

where the last equality follows from the LLN and the induction hypothesis, the series of inequalities follow from repeated use of the Cauchy-Schwartz inequality and the fact that $\big|\max_z f(z) - \max_z g(z)\big| \leq \max_z \big|f(z) - g(z)\big|$. This proves the first part of the result. The second part of the result follows from an identical argument since $\sqrt{n}(\hat{\beta}_T^{(b)} - \hat{\beta}_T)$ converges to the same limiting distribution as $\sqrt{n}(\hat{\beta}_T - \beta_T^{*})$ in probability by the bootstrap central limit theorem (see for example Bickel and Freedman 1981) and hence satisfies the condition stated in the theorem. The same induction argument succeeds with only minor changes in notation. □

Lemma 3.8.3. *Assume (A1)-(A3). Define* $\Delta_t(h_{t,1}) \triangleq \# \left\{ \hat{\mathcal{A}}_t(h_{t,1}) \Delta \mathcal{A}_t^{*}(h_{t,1}) \right\}$. *Let* $\epsilon > 0$ *be arbitrary. There exists subset* $B_{t,\epsilon}$ *of* $\mathbb{R}^{\dim(\beta_{t,1}^{*})}$ *satisfying* $P(H_{t,1} \in B_{t,\epsilon}) \geq 1 - \epsilon$, *and* $\sup_{h_{t,1} \in B_{t,\epsilon}} \Delta_t(h_{t,1}) = o_P(1)$.

Proof. If $h_{t,1}$ is any patient history for which $\#\mathcal{A}_t^{*}(h_{t,1}) = 1$ then $\Delta_t(h_{t,1}) 1_{\#\mathcal{A}_t^{*}(h_{t,1}) > 1} \equiv 0$. Let $h_{t,1}$ be a patient history for which $\mathcal{A}_t^{*}(h_{t,1}) > 1$ and let i denote an arbitrary element in $\mathcal{A}_t^{*}(h_{t,1})$. Then,

$$
\begin{aligned}
T_{t,n,i}(h_{t,1})/\lambda_n &\triangleq \frac{\left(h_{t,1}^{\intercal} \left(\mathbb{V}_{t,n,i} + \sqrt{n} \beta_{t,i}^{*} \right) - \max_{j \neq i} h_{t,1}^{\intercal} \left(\mathbb{V}_{t,n,j} + \sqrt{n} \beta_{t,j}^{*} \right) \right)^2}{\lambda_n h_{t,1}^{\intercal} \hat{\zeta}_{t,i} h_{t,1}} \\
&\leq \frac{\left(h_{t,1}^{\intercal} \mathbb{V}_{t,n,i} - \max_{j \neq i} h_{t,1}^{\intercal} \left(\mathbb{V}_{t,n,j} + \sqrt{n} (\beta_{t,j}^{*} - \beta_{t,i}^{*}) \right) \right)^2}{\lambda_n ||h_{t,1}||^2 \sigma_{t,n,i}}, \qquad (3.51)
\end{aligned}
$$

where $\sigma_{t,n,i}$ is the smallest eigenvalue of $\hat{\zeta}_{t,i}$. Using the triangle inequality, the last term on the right hand side (3.51) is further bounded above by

$$\frac{\left(\left|h_{t,1}^{\intercal}\mathbb{V}_{t,n,i}\right| + \left|\max_{j\neq i} h_{t,1}^{\intercal}\left(\mathbb{V}_{t,n,j} + \sqrt{n}(\beta_{t,j}^{*} - \beta_{t,i}^{*})\right|\right)\right)^{2}}{\lambda_n ||h_{t,1}||^2 \sigma_{t,n,i}}.$$

Next we note that

$$\frac{\left|\max_{j\neq i} h_{t,1}^{\intercal}\left(\mathbb{V}_{t,n,j} + \sqrt{n}(\beta_{t,j}^{*} - \beta_{t,i}^{*})\right)\right|}{||h_{t,1}||} \leq \max_{j\neq i} ||\mathbb{V}_{t,n,j}||_{\infty}.$$

To see this, let $a(i) \triangleq \arg\max_{j\neq i} h_{t,1}^{\intercal}\left(\mathbb{V}_{t,n,j} + \sqrt{n}(\beta_{t,j}^{*} - \beta_{t,i}^{*})\right)$, and let $i' \neq i$ belong to $\mathcal{A}_t^{*}(h_{t,1})$ and notice that

$$h_{t,1}^{\intercal}\mathbb{V}_{t,n,i'} \leq h_{t,1}^{\intercal}\left(\mathbb{V}_{t,n,a(i)} + \sqrt{n}(\beta_{t,a(i)}^{*} - \beta_{t,i}^{*})\right) \leq h_{t,1}^{\intercal}\mathbb{V}_{t,n,a(i)}.$$

The above set of inequalities make use of the fact that $h_{t,1}^{\intercal}(\beta_{t,j}^{*} - \beta_{t,i}^{*}) \leq 0$ with equality holding when $j \in \mathcal{A}_t^{*}(h_{t,1})$. This result, the preceding discussion, and some algebra show that $T_{t,n,i}(h_{t,1})/\lambda_n$ is bounded above by

$$\frac{\left(||\mathbb{V}_{t,n,i}||_{\infty} + \max_{j\neq i} ||\mathbb{V}_{t,n,j}||_{\infty}\right)^2}{\lambda_n \min_{1\leq i\leq K_t} \sigma_{t,n,i}} \leq \frac{4\max_{1\leq j\leq K_t} ||\mathbb{V}_{t,n,j}||_{\infty}^2}{\lambda_n \min_{1\leq i\leq K_t} \sigma_{t,n,i}}.$$

The right hand side of the above display is $o_P(1)$ and depends neither on $h_{t,1}$ nor the choice of $i \in \mathcal{A}_t^{*}(h_{t,1})$. We have shown that

$$\sup_{h_{t,1}\in\mathbb{R}^{\dim(\beta_{t,1}^{*})}} \#(\mathcal{A}_t^{*}(h_{t,1}) \setminus \hat{\mathcal{A}}_t(h_{t,1}))1_{\#\mathcal{A}_t^{*}(h_{t,1})>1} = o_P(1).$$

Now let $\epsilon > 0$ be arbitrary, then for each $i = 1, 2, \ldots, K_t$ we can choose $\eta_i > 0$

sufficiently small so that

$$P\left(0 < \frac{\left|H_{t,1}^{\intercal}\beta_{t,i}^{*} - \max_{j\neq i} H_{t,1}^{\intercal}\beta_{t,j}^{*}\right|}{||H_{t,1}||} < \eta_i\right) \leq \epsilon/K_t,$$

where we have defined $0/0 = 0$ for convenience. Let $\eta \triangleq \min_{1\leq i\leq K_t}\eta_i$ and define

$$B_{t,\epsilon,i} \triangleq \left\{\frac{\left|H_{t,1}^{\intercal}\beta_{t,i}^{*} - \max_{j\neq i} H_{t,1}^{\intercal}\beta_{t,j}^{*}\right|}{||H_{t,1}||} = 0\right\}\bigcup\left\{\frac{\left|H_{t,1}^{\intercal}\beta_{t,i}^{*} - \max_{j\neq i} H_{t,1}^{\intercal}\beta_{t,j}^{*}\right|}{||H_{t,1}||} > \eta\right\},$$

so that $P(H_{t,1} \in B_{t,\epsilon,i}) \geq 1 - \epsilon/K_t$. Define $B_{t,\epsilon} \triangleq \bigcap_{i=1}^{K_t} B_{t,\epsilon,i}$, the union bound ensures that $P(H_{t,1} \in B_{t,\epsilon}) \geq 1 - \epsilon$. Let $h_{t,1}$ be arbitrary element of $B_{t,\epsilon}$. We now consider the limiting behavior of $T_{t,n,i}(h_{t,1})/\lambda_n$. We consider the following three cases: (i) $i \in \mathcal{A}_t^*(h_{t,1})$ and $\#\mathcal{A}_t^*(h_{t,1}) = 1$, (ii) $i \in \mathcal{A}_t^*(h_{t,1})$ and $\#\mathcal{A}_t^*(h_{t,1}) > 1$, and (iii) $i \notin \mathcal{A}_t^*(h_{t,1})$. In the first case, $\sqrt{T_{t,n,i}(h_{t,1})/\lambda_n}$ is bounded below by

$$\frac{\left|h_{t,1}^{\intercal}\mathbb{V}_{t,n,i} - \max_{j\neq i} h_{t,1}^{\intercal}\left(\mathbb{V}_{t,n,j} + \sqrt{n}(\beta_{t,j}^{*} - \beta_{t,i}^{*})\right)\right|}{\sqrt{\lambda_n \nu_{t,n,i}}||h_{t,1}||},$$

where $\nu_{t,n,i}$ denotes the largest eigenvalue of $\hat{\zeta}_i$. We can further bound the above quantity by

$$\min_{j\neq i}\frac{\left|h_{t,1}^{\intercal}\mathbb{V}_{t,n,i} - h_{t,1}^{\intercal}\left(\mathbb{V}_{t,n,j} + \sqrt{n}(\beta_{t,j}^{*} - \beta_{t,i}^{*})\right)\right|}{\sqrt{\lambda_n \nu_{t,n,i}}||h_{t,1}||},$$

which, in turn, is further bounded below by

$$\frac{\min_{j\neq i}\left|h_{t,1}^{\intercal}\sqrt{n}(\beta_{t,j}^{*} - \beta_{t,i}^{*})\right| - 2\max_{1\leq i\leq K_t}\left|h_{t,1}^{\intercal}\mathbb{V}_{t,n,i}\right|}{\sqrt{\lambda_n \nu_{t,n,i}}||h_{t,1}||}.$$

A final lower bound is obtained by taking the infimum over all $h_{t,1}$ in $B_{t,\epsilon}$ and the maximum over the eigenvalues $\nu_{t,n,i}$. This bound is given by

$$\frac{\sqrt{n}\eta - 2\max_{1\leq i\leq K_t}||\mathbb{V}_{t,n,i}||_\infty}{\sqrt{\lambda_n \max_{1\leq i\leq K_t}\nu_{t,n,i}}},$$

which does not depend on either $h_{t,1}$ or i, and diverges to ∞ in probability as n tends to ∞. Consider next case (ii). It was shown above that if $h_{t,1}$ is such that $\#\mathcal{A}_t^*(h_{t,1}) > 1$ and $i \in \mathcal{A}_t^*(h_{t,1})$ then $T_{t,n,i}(h_{t,1})/\lambda_n$ is bounded above by

$$\frac{\left(||\mathbb{V}_{t,n,i}||_\infty + \max_{j\neq i}||\mathbb{V}_{t,n,j}||_\infty\right)^2}{\lambda_n \min_{1\leq i\leq K_t}\sigma_{t,n,i}} \leq \frac{4\max_{1\leq j\leq K_t}||\mathbb{V}_{t,n,j}||_\infty^2}{\lambda_n \min_{1\leq i\leq K_t}\sigma_{t,n,i}},$$

which does not depend on either $h_{t,1}$ or i and converges to zero in probability as n tends to ∞. To address case (iii) suppose that $h_{t,1}$ belongs to $B_{t,\epsilon}$ and that $i \notin \mathcal{A}_t^*(h_{t,1})$. Following the arguments given above, $\sqrt{T_{t,n,i}(h_{t,1})/\lambda_n}$ is bounded below by

$$\frac{\left|h_{t,1}^{\intercal}\mathbb{V}_{t,n,i} - \max_{j\neq i} h_{t,1}^{\intercal}\left(\mathbb{V}_{t,n,j} + \sqrt{n}(\beta_{t,j}^* - \beta_{t,i}^*)\right)\right|}{\sqrt{\lambda_n \min_{1\leq i\leq K_t}\nu_{t,n,i}}||h_{t,1}||}.$$

A further lower bound is given by

$$\frac{\left|\max_{j\neq i} h_{t,1}^{\intercal}(\mathbb{V}_{t,n,j} + \sqrt{n}(\beta_{t,j}^* - \beta_{t,i}^*)\right| - \left|h_{t,1}^{\intercal}\mathbb{V}_{t,n,i}\right|}{\sqrt{\lambda_n \min_{1\leq i\leq K_t}\nu_{t,n,i}}||h_{t,1}||} \geq \frac{\sqrt{n}\eta - 2\max_{1\leq i\leq K_t}||\mathbb{V}_{t,n,i}||_\infty}{\sqrt{\lambda_n \min_{1\leq i\leq K_t}\nu_{t,n,i}}}.$$

The right hand side of the above display diverges to ∞ in probability as n tends to ∞. Furthermore, this lower bound does not depend on either $h_{t,1}$ or i. The results for cases (i) and (ii) show that $\sup_{h_{t,1}\in B_{t,\epsilon}} \#\left(\mathcal{A}_t^*(h_{t,1}) \setminus \hat{\mathcal{A}}_t(h_{t,1})\right) = o_P(1)$. The result for case (ii) shows that $\sup_{h_{t,1}\in B_{t,\epsilon}} \#\left(\hat{\mathcal{A}}_t(h_{t,1}) \setminus \mathcal{A}_t^*(h_{t,1})\right) = o_p(1)$. □

Corollary 3.8.4. *Assume (A1)-(A3). Then for each $t \geq 2$*

$$\mathbb{P}_n||B_{t-1}||\,||H_{t,1}||1_{\Delta_t(H_{t,1})>0} = O_P(1).$$

Proof. Let $\eta > 0$ be arbitrary, we intend to show that

$$P\left(\mathbb{P}_n||B_{t-1}||\,||H_{t,1}||1_{\Delta_t(H_{t,1})>0} > \eta\right) \to 0$$

as n tends to ∞. Choose ϵ sufficiently small so that $P||B_{t-1}||\,||H_{t,1}||1_{H_{t,1}\notin B_{t,\epsilon}} \leq \eta/2$. where $B_{t,\epsilon}$ is as defined in Lemma (3.8.3). Then notice that

$$\mathbb{P}_n||B_{t-1}||\,||H_{t,1}||1_{\Delta_t(H_{t,1})>0} \leq \mathbb{P}_n||B_{t-1}||\,||H_{t,1}|| \sup_{h_{t,1}\in B_{t,\epsilon}} 1_{\Delta_t(h_{t,1})>0} + \mathbb{P}_n||B_{t-1}||\,||H_{t,1}||1_{h_{t,1}\notin B_{t,\epsilon}}.$$

The right hand side of the above display is smaller than η with probability tending to one by appeal to Lemma (3.8.3), the LLN, and Slutsky's theorem. □

Proof of supplement Theorem 1.1 part 1. This proof proceeds by induction on t and for convenience is broken into several intermediate steps.

Step 1. In this step we establish the limiting distribution of $\sqrt{n}(\hat{\beta}_{T-1} - \beta^*_{T-1})$. Standard results for least squares show that $\sqrt{n}(\hat{\beta}_T - \beta^*_T)$ is asymptotically normal with mean zero. Recall that $\sqrt{n}(\hat{\beta}_{T-1} - \beta^*_{T-1})$ is equal to

$$\mathbb{W}'_{T-1,n} + \Sigma^{-1}_{T-1,n}\mathbb{P}_n B^{\intercal}_{T-1}\sqrt{n}(\hat{\beta}_{T-1,0} - \beta^*_{T-1,0}) + \Sigma^{-1}_{T-1,n}\mathbb{P}_n B^{\intercal}_{T-1}\mathbb{U}_{T,n}, \tag{3.52}$$

where

$$\mathbb{U}_{T,n} \triangleq \sqrt{n}\left(\max_{1\leq i\leq K_T} H^{\intercal}_{T,1}\hat{\beta}_{T,i} - \max_{1\leq i\leq K_T} H^{\intercal}_{T,1}\beta^*_{T,i}\right).$$

Notice that $||\mathbb{P}_n B^{\intercal}_1\mathbb{U}_{T,n}||$ is $O_p(1)$ by appeal to Lemma (3.8.1), the asymptotic normality of $\sqrt{n}(\hat{\beta}_T - \beta_T^*)$,and (A1). Thus, by Slutsky's theorem we can replace $\Sigma_{T-1,n}^{-1}$ in (3.52) with $\Sigma_{T-1,\infty}^{-1}$ whilst incurring an error of no more than $o_P(1)$. Let $c \in \mathbb{R}^{\dim(B_1)}$ and $M > 0$ be arbitrary. The class of functions

$$\mathcal{F} \triangleq \left\{ f(\beta_T) = c^\intercal \Sigma_{T-1,\infty}^{-1} B_{T-1}^\intercal \left(\max_{1 \le i \le K_T} H_{T,1}^\intercal \beta_{T,i} \right) \; : \; \beta_T \in \mathcal{B}(M, \beta_T^*) \right\}$$

is a P-measurable BUEI class with square integrable envelope $||\Sigma_{T-1,\infty}^{-1} B_{T-1}^\intercal|| ||H_{T-1}||$ and hence Donsker. This fact is established using BUEI and measurability preservation properties (see Lemma 9.16 part (iii) in Kosorok 2008). Using the consistency of $\hat{\beta}_T$ and stochastic equicontinuity, it follows that

$$\begin{aligned} \Sigma_{T-1,\infty}^{-1} \mathbb{P}_n B_{T-1}^\intercal \mathbb{U}_{T,n} &= \Sigma_{T-1,\infty}^{-1} P B_{T-1}^\intercal \mathbb{U}_{T,n} + \Sigma_{T-1,\infty}^{-1} (\mathbb{P}_n - P) B_{T-1}^\intercal \mathbb{U}_{T,n} \\ &= \Sigma_{T-1,\infty}^{-1} P B_{T-1}^\intercal \mathbb{U}_{T,n} + o_P(1). \end{aligned}$$

In particular, let $\mathbb{G}_n \triangleq \sqrt{n}(\mathbb{P}_n - P)$ then $c^\intercal \Sigma_{T-1,\infty}^{-1} (\mathbb{P}_n - P) B_{T-1}^\intercal \mathbb{U}_{T,n} = \mathbb{G}_n \left(f(\hat{\beta}_T) - f(\beta_T^*) \right)$. For $\delta, \epsilon > 0$ arbitrary we can choose M sufficiently large so that $P\left(||\hat{\beta}_T - \beta_T^*|| > M/\sqrt{n} \right) \le \delta/2$. Thus,

$$\begin{aligned} P\left(\left| G_n \left(f(\hat{\beta}_T) - f(\beta_T^*) \right) \right| > \epsilon \right) &\le P\left(||\hat{\beta}_T - \beta_T^*|| > M/\sqrt{n} \right) \\ &\quad + P\left(\sup_{s \in \mathcal{B}(M/\sqrt{n}, \beta_T^*)} \left| \mathbb{G}_n \left(f(s) - f(\beta_T^*) \right) \right| > \epsilon \right). \end{aligned}$$

The right hand side of the above display can be made to be less than δ for sufficiently large n by appeal to the equicontinuity of $\mathbb{G}_n$. Finally, since c was arbitrary we have shown convergence in probability to zero of each component $\Sigma_{T-1,\infty}^{-1} (\mathbb{P}_n - P) B_{T-1}^\intercal \mathbb{U}_{T,n}$ and thus convergence in Euclidean norm follows.

Let $\Delta_T(H_{T,1})$ is as defined in Lemma (3.8.3) As a next step write

$$\begin{aligned}\Sigma_{T-1,\infty}^{-1} PB_{T-1}^{\intercal}\mathbb{U}_{T,n} &= \Sigma_{T-1,\infty}^{-1} PB_{T-1}^{\intercal} \max_{i\in\mathcal{A}_T^*(H_{T,1})} H_{T,1}^{\intercal}\mathbb{V}_{T,n,i} 1_{\Delta_T(H_{T,1})=0} + \Sigma_{T-1,\infty}^{-1} PB_{T-1}^{\intercal}\mathbb{U}_{T,n} 1_{\Delta_T(H_{T,1})} \\ &= \Sigma_{T-1,\infty}^{-1} PB_{T-1}^{\intercal} \max_{i\in\mathcal{A}_T^*(H_{T,1})} H_{T,1}^{\intercal}\mathbb{V}_{T,n,i} + o_P(1),\end{aligned}$$

where the last equality follows from Lemma (3.8.3), Lemma (3.8.1), and (A1). The map $\delta \mapsto \Sigma_{T-1,\infty}^{-1} PB_{T-1}^{\intercal} \max_{i\in\mathcal{A}_T^*(H_{T,1})} H_{T,1}^{\intercal}\delta_i$ is continuous with respect to the usual metric on $\mathbb{R}^{N_T}$. The remainder of the derivation of the limiting distribution of $\mathbb{V}_{T-1,n} = \sqrt{n}(\hat{\beta}_{T-1} - \beta_{T-1}^*)$ follows from the continuous mapping theorem and the central limit theorem.

Step 2. In this step we derive the limiting distribution of $\sqrt{n}(\hat{\beta}_{T-2} - \beta_{T-2}^*)$. Which can be decomposed as

$$\mathbb{W}'_{T-2,n} + \Sigma_{T-2,n}^{-1}\mathbb{P}_n B_{T-2}^{\intercal} H_{T-1,0}^{\intercal}\sqrt{n}(\hat{\beta}_{T-1,0} - \beta_{T-1,0}^*) + \Sigma_{T-2,n}^{-1}\mathbb{P}_n B_{T-2}^{\intercal}\mathbb{U}_{T-1,n},$$

where

$$\mathbb{U}_{T-1,n} \triangleq \sqrt{n}\left(\max_{1\leq i\leq K_{T-1}} H_{T-1,1}^{\intercal}\hat{\beta}_{T-1,i} - \max_{1\leq i\leq K_{T-1}} H_{T,1}^{\intercal}\beta_{T-1,i}^*\right)$$

and

$$\begin{aligned}\mathbb{W}'_{T-2,n} &= \Sigma_{T-2,n}^{-1}\sqrt{n}\mathbb{P}_n B_{T-2}^{\intercal}\left(Y_{T-2} + H_{T-1,0}^{\intercal}\beta_{T-1,0}^* + \max_{1\leq i\leq K_{T-1}} H_{T-1,1}^{\intercal}\beta_{T-1,i}^* - B_{T-2}^{\intercal}\beta_{T-2}^*\right) \\ &= \Sigma_{T-2,n}^{-1}\sqrt{n}(\mathbb{P}_n - P)B_{T-2}^{\intercal}\left(Y_{T-2} + H_{T-1,0}^{\intercal}\beta_{T-1,0}^* + \max_{1\leq i\leq K_{T-1}} H_{T-1,1}^{\intercal}\beta_{T-1,i}^* - B_{T-2}^{\intercal}\beta_{T-2}^*\right).\end{aligned}$$

Thus, $\mathbb{W}'_{T-2,n}$ is seen to be asymptotically normal from the central limit theorem and Slutsky's theorem. Next, we note that the same arguments supplied in step 1 above can be used to show that

$$\begin{aligned}\mathbb{P}_n B_{T-2}^{\intercal}\mathbb{U}_{T-1,n} &= PB_{T-2}^{\intercal}\mathbb{U}_{T-1,n} + (\mathbb{P}_n - P)B_{T-2}^{\intercal}\mathbb{U}_{T-1,n} \\ &= PB_{T-2}^{\intercal}\mathbb{U}_{T-1,n} + o_P(1).\end{aligned}$$

Furthermore, it follows that

$$\begin{aligned}PB_{T-2}^{\intercal}\mathbb{U}_{T-1,n} &= PB_{T-2}^{\intercal}\sqrt{n}\left(\max_{1\le i\le K_{T-1}} H_{T-1}^{\intercal}\hat{\beta}_{T-1,i} - \max_{1\le i\le K_{T-1}} H_{T-1,1}^{\intercal}\beta_{T-1,i}^{*}\right) \\ &= PB_{T-2}^{\intercal}\sqrt{n}\left(\max_{i\in\hat{\mathcal{A}}_{T-1}(H_{T-1,1})} H_{T-1,1}^{\intercal}\hat{\beta}_{T-1,i} - \max_{i\in\mathcal{A}_{T-1}^{*}(H_{T-1,1})} H_{T-1,1}^{\intercal}\beta_{T-1,i}^{*}\right)\end{aligned}$$

which can be decomposed as

$$\begin{aligned}&PB_{T-2}^{\intercal}\max_{i\in\mathcal{A}_{T-1}^{*}(H_{T-1,1})} H_{T-1,1}^{\intercal}\sqrt{n}(\hat{\beta}_{T-1,i} - \beta_{T-1,i})1_{\Delta_{T-1}(H_{T-1,1})=0} \\ &+PB_{T-2}^{\intercal}\sqrt{n}\left(\max_{i\in\hat{\mathcal{A}}_{T-1}(H_{T-1,1})} H_{T-1,1}^{\intercal}\hat{\beta}_{T-1,i} - \max_{i\in\mathcal{A}_{T-1}^{*}(H_{T-1,1})} H_{T-1,1}^{\intercal}\beta_{T-1,i}^{*}\right)1_{\Delta_{T-1}(H_{T-1,1})>0}.\end{aligned}$$

Using Lemma (3.8.3) and following the same argument as given in step 1, it can be concluded that the quantity in the above display is equal to

$$PB_{T-2}^{\intercal}\max_{i\in\mathcal{A}_{T-1}^{*}(H_{T-1,1})} H_{T-1,1}^{\intercal}\sqrt{n}(\hat{\beta}_{T-1,i} - \beta_{T-1,i}^{*}) + o_P(1).$$

The map $\delta \mapsto PB_{T-2}^{\intercal}\max_{i\in\mathcal{A}_{T-1}^{*}(H_{T-1,1})} H_{T-1,1}^{\intercal}\delta_i$ is continuous on $\mathbb{R}^{N_{T-1}}$. Consequently, the continuous mapping theorem shows that $c^{\intercal}\sqrt{n}(\hat{\beta}_{T-1} - \beta_{T-1}^{*})$ is equal to

$$c^{\intercal}\mathbb{W}_{T-2,n}' + c^{\intercal}\Sigma_{T-2,n}^{\intercal}PB_{T-2}^{\intercal}\sqrt{n}(\hat{\beta}_{T-1,0} - \beta_{T-1,0}^{*})$$

$$+ c^{\intercal}\Sigma_{T-2,n}^{-1} P B_{T-2}^{\intercal} \max_{i\in\mathcal{A}_{T-2}^{*}(H_{T-1,1})} H_{T-1,1}^{\intercal}\sqrt{n}(\hat{\beta}_{T-1,i} - \beta_{T-1,i}^{*}) + o_P(1),$$

which converges to the desired limit by the central limit theorem and Slutsky's theorem.

Step 3. The preceding argument used to establish the limiting distribution of $c^{\intercal}\sqrt{n}(\hat{\beta}_{T-2} - \beta_{T-2}^{*})$ depended on the preceding stage only through $\sqrt{n}(\hat{\beta}_{T-1} - \beta_{T-1}^{*})$ and the sets $\hat{\mathcal{A}}_{T-1}(h_{T-1,1})$ and $\mathcal{A}_{T-1}^{*}(h_{T-1,1})$. Thus, since the limiting behavior of $\hat{\mathcal{A}}_t(h_{t,1})$ has been described for all t in Lemma 3.8.3, the result for all t follows from backwards induction on t. □

Lemma 3.8.5. *Assume (A1) and (A2). Let $\epsilon > 0$ and $K > 0$ be arbitrary. Fix a vector $c \in \mathbb{R}^{\dim(B_{T-1})}$ and define, for each $\delta \in \mathcal{B}(K, 0) \subset \mathbb{R}^{N_T}$, a function in $l^{\infty}(\mathbb{R}^{N_T})$ given by*

$$g_{T,\delta}(\gamma) \triangleq P c^{\intercal} B_{T-1}^{\intercal}\left(\max_{i\in\mathcal{A}_T^{*}(H_{T,1})} H_{T,1}^{\intercal}(\delta_i + \gamma_i) - \max_{i\in\mathcal{A}_T^{*}(H_{T,1})} H_{T,1}^{\intercal}\gamma_i\right) 1_{\#\mathcal{A}_T^{*}(H_{T,1})>1}.$$

Then, there exists $M > 0$ so that

$$\sup_{\delta\in\mathcal{B}(K,0)}\left|\sup_{\delta\in\mathcal{B}(M,0)} g_{T,\delta}(\gamma) - \sup_{\gamma\in\mathbb{R}^{N_T}} g_{T,\delta}(\gamma)\right| < \epsilon.$$

Proof. See the proof of Lemma (3.6.10). □

Lemma 3.8.6. *Assume (A1) and (A2). Let $M > 0$ be arbitrary and let $g_{T,\delta}$ be as defined in Lemma (3.8.5). For an arbitrary subset V of $\mathbb{R}^{\dim(\beta_{T,1}^{*})}$, and fixed value $\delta \in \mathbb{R}^{\dim(\beta_{T,1}^{*})}$, let $g_{T,\delta}(V)$ denote the image of V under the map $g_{T,\delta}$. Then, for all sufficiently large n $g_{T,\delta}(\mathcal{B}(M, 0)) \subseteq g_{T,\delta}(\mathcal{S}_{T,n})$.*

Proof. Let $\nu \in \mathcal{B}(M, 0)$ be arbitrary and let ν' satisfy $\nu_i' = \nu_i + \sqrt{n}\beta_{T,i}^{*}$ for all $i = 1, 2, \ldots, K_T$, then $g_{T,\delta}(\nu) = g_{T,\delta}(\nu')$. Furthermore, $||\nu' - \sqrt{n}\beta_{T,:}^{*}||_{\infty}^{2} = ||\nu||_{\infty}^{2} \leq \sqrt{N_T}M$

which is smaller than λ_n for sufficiently large n. This proves the result. □

Corollary 3.8.7. *Assume (A1) and (A2). Let $K > 0$ and $M > 0$ be arbitrary ,and fix a vector $c \in \mathbb{R}^{\dim(B_{T-1})}$. For each $\delta \in \mathcal{B}(K, 0) \subset \mathbb{R}^{N_T}$ let $g_{T,\delta}$ be as defined in the Lemma (3.8.5). Then,*

$$\inf_{\delta \in \mathcal{B}(K,0)} \left(\sup_{\gamma \in \mathcal{S}_{T,n}} g_{T,\delta}(\gamma) - \sup_{\gamma \in \mathcal{B}(M,0)} g_{T,\delta}(\gamma) \right) \geq 0.$$

Proof. Let n be sufficiently large so that $\lambda_n > \sqrt{N_T} M$. Then, by appeal to the preceding lemma, it follows that for arbitrary $\delta \in \mathbb{R}^{\dim(\beta^*_{T,1})}$, $\sup_{\gamma \in \mathcal{S}_{T,n}} g_{T,\delta}(\gamma) - \sup_{\gamma \in \mathcal{B}(M,0)} g_{T,\delta}(\gamma) \geq 0$. Take the infimum over δ to obtain the result. □

Proof of supplement Theorem 1.1 part 2. We will derive the limiting distribution of $\mathbb{V}_{T-1}(\gamma_T)$ as a process in $l^\infty(\mathbb{R}^{N_T})$. Recall that $\mathbb{V}_{T-1,n}(\gamma_T)$ has the following decomposition

$$\begin{aligned} \mathbb{V}_{T-1,n}(\gamma_T) &= \mathbb{W}'_{T-1,n} + \Sigma^{-1}_{T-1,n} \mathbb{P}_n B^\intercal_{T-1} H^\intercal_{T,0} \mathbb{V}_{T,n,0} + \Sigma^{-1}_{T-1,n} \mathbb{P}_n B^\intercal_{T-1} \mathbb{U}_{T,n} 1_{\#\hat{\mathcal{A}}_T(H_{T,1})=1} \\ &+ \Sigma^{-1}_{T-1,n} \mathbb{P}_n B^\intercal_{T-1} \left[\max_{i \in \tilde{\mathcal{A}}_T(H_{T,1})} H^\intercal_{T,1}(\mathbb{V}_{T,n,i} + \gamma_{T,i}) - \max_{i \in \tilde{\mathcal{A}}_T(H_{T,1})} H^\intercal_{T,1} \mathbb{V}_{T,n,i} \right] 1_{\#\hat{\mathcal{A}}_T(H_{T,1})>1}. \end{aligned} \tag{3.53}$$

The limiting behavior of the first three terms in (3.53) was described in the proof of Theorem 1.1 part 1. Thus, we only consider the behavior of the last term of (3.53). Using Lemma (3.8.1) it is seen that

$$\left|\left| \mathbb{P}_n B_{T-1} \left[\max_{i \in \tilde{\mathcal{A}}_T(H_{T,1})} H^\intercal_{T,1}(\mathbb{V}_{T,n,i} + \gamma_{T,i}) - \max_{i \in \tilde{\mathcal{A}}_T(H_{T,1})} H^\intercal_{T,1} \mathbb{V}_{T,n,i} \right] 1_{\#\hat{\mathcal{A}}_T(H_{T,1})>1} \right|\right|$$

is bounded above by $\mathbb{P}_n ||B_{T-1}||\, ||H_{T,1}|| \max_{1 \leq i \leq K_T} ||\mathbb{V}_{T,n,i}|| = O_p(1)$. Thus, we can replace $\Sigma_{T-1,n}^{-1}$ with $\Sigma_{T-1,\infty}^{-1}$ by appeal to Slutsky's theorem. Furthermore, use of Lemma (3.8.1) and Corollary (3.8.4) shows that

$$\Sigma_{T-1,\infty}^{-1}\mathbb{P}_n B_{T-1}\left[\max_{i\in\tilde{\mathcal{A}}_T(H_{T,1})} H_{T,1}^{\intercal}(\mathbb{V}_{T,n,i}+\gamma_{T,i}) - \max_{i\in\tilde{\mathcal{A}}_T(H_{T,1})} H_{T,1}^{\intercal}\mathbb{V}_{T,n,i}\right]1_{\#\hat{\mathcal{A}}_T(H_{T,1})>1}$$

can be replaced with

$$\Sigma_{T-1,\infty}^{-1}\mathbb{P}_n B_{T-1}\left[\max_{i\in\mathcal{A}^*_T(H_{T,1})} H_{T,1}^{\intercal}(\mathbb{V}_{T,n,i}+\gamma_{T,i}) - \max_{i\in\mathcal{A}^*_T(H_{T,1})} H_{T,1}^{\intercal}\mathbb{V}_{T,n,i}\right]1_{\#\mathcal{A}^*_T(H_{T,1})>1}, \tag{3.54}$$

up to a term of order $o_P(1)$. Next, notice that the class of functions given by

$$\mathcal{F} \triangleq \left\{\max_{i\in\mathcal{A}^*_T(H_{T,1})} H_{T,1}^{\intercal}(\delta_i+\gamma_{T,i}) - \max_{i\in\mathcal{A}^*_T(H_{T,1})} H_{T,1}^{\intercal}\gamma_{T,i}\,,\ \delta\in\mathcal{B}(K,0)\,\gamma\in\mathbb{R}^{N_T}\right\}$$

is a P-measurable BUEI class with square integrable envelope and hence is Donsker. One way to see this is to reparametrize members of $\mathcal{F}$ as

$$\begin{aligned}\max_{1\le i\le K_T}&\left(H_{T,1}^{\intercal}(\delta_i+\gamma_{T,i})1_{i\in\mathcal{A}^*_T(H_{T,1})} - \sum_{i=1}^{K_T}\left(1+|H_{T,1}^{\intercal}(\delta_i+\gamma_{T,i})|\right)1_{i\notin\mathcal{A}^*_T(H_{T,1})}\right)\\ &\qquad - \max_{1\le i\le K_T}\left(H_{T,1}^{\intercal}\gamma_{T,i}1_{i\in\mathcal{A}^*_T(H_{T,1})} - \left[\sum_{j=1}^{K_T}\left(1+|H_{T,1}^{\intercal}\gamma_{T,j}|\right)\right]1_{i\notin\mathcal{A}^*_T(H_{T,1})}\right)\end{aligned}$$

and notice that this class can be constructed by applying pointwise maximizations and additions of VC classes. Since these operations are VC preserving (see for example, Lemma 9.9 Kosorok 2008) and $\mathcal{F}$ has envelope $||H_{T,1}||K$ which is square integrable, the class $\mathcal{F}$ is BUEI. Measurability is seen by arguing that the class can be arbitrarily closely approximated by a class that restricts γ and δ be rational vectors. Finally, it follows that for any unit vector $c\in\mathbb{R}^{\dim(B_1)}$ the class $c^{\intercal}\Sigma_{T-1,\infty}^{-1}B_{T-1}\mathcal{F}$ is also P-measurable BUEI with square integrable envelope proportional to $||B_{T-1}||\,||H_{T,1}||$.

Thus, we can replace $\mathbb{P}_n$ with P in (3.54) up to a term of order $o_P(1)$. The final desired result follows from the continuous mapping theorem. □

Proof of supplement Theorem 1.1 part 3. The proof proceeds by induction. The limiting distribution of $\mathbb{V}_{T-1,n}(\gamma_T)$ was derived in the proof of Supplement Theorem 1.1 part 2. We now consider the limiting distribution of $\mathbb{V}_{T-2,n}((\gamma_{T-1}, \gamma_{T-2}))$. Recall that

$$
\begin{aligned}
\mathbb{V}_{T-2,n}((\gamma_{T-1}, \gamma_{T-1})) &= \mathbb{W}'_{T-2,n} + \Sigma^{-1}_{T-2,n}\mathbb{P}_n B^\intercal_{T-2} H^\intercal_{T-1,1}\mathbb{V}_{T-1,n,0}(\gamma_T) \\
&+ \Sigma^{-1}_{T-2,n}\mathbb{P}_n B^\intercal_{T-2}\left(\mathbb{U}_{T-2,n} - \max_{i\in\tilde{\mathcal{A}}_{T-1}(H_{T-1,1})} H^\intercal_{T-1,1}\mathbb{V}_{T-1,n,i}\right)1_{\#\hat{\mathcal{A}}_{T-1}(H_{T-1,1})=1} \\
+\Sigma^{-1}_{T-2,n}\mathbb{P}_n B^\intercal_{T-2}\Bigg(&\max_{i\in\tilde{\mathcal{A}}_{T-1}(H_{T-1,1})} H^\intercal_{T-1,1}\left(\mathbb{V}_{T-1,n,i}(\gamma_T) + \gamma_{T-1,i}\right) - \max_{i\in\tilde{\mathcal{A}}_{T-1}(H_{T-1,1})} H^\intercal_{T,1}\gamma_{T-1,i} \\
&- \max_{i\in\tilde{\mathcal{A}}_{T-1}(H_{T-1,1})} H^\intercal_{T,1}\mathbb{V}_{T-1,n,i}(\gamma_T)\Bigg)1_{\#\hat{\mathcal{A}}_{T-1}(H_{T-1,1})>1}. \qquad (3.55)
\end{aligned}
$$

The limiting distribution of $\mathbb{W}'_{T-2}$ was derived in the proof of Supplement Theorem 1.1 part 1. The limiting distribution of the second term in (3.55) follows from the continuous mapping theorem and the LLN. The third term in (3.55) is seen to be $o_P(1)$ by first noting that this term is zero whenever $\hat{\mathcal{A}}_{T-1}(H_{T-1,1}) = \mathcal{A}^*_{T-1}(H_{T-1,1})$ and then applying Lemma (3.8.3) and arguments similar to those given in parts 1 and 2 of Supplement Theorem 1.1. We will focus attention on the last term in (3.55). As a first step, we note that this term is seen to be $O_P(1)$ by Lemma (3.8.1) and Supplement Theorem 1.1 part 2. Thus, by appeal to Slutsky's theorem, we can replace $\Sigma^{-1}_{T-2,n}$ with $\Sigma^{-1}_{T-2,\infty}$. Arguments paralleling those given in the proof of Supplement Theorem 1.1 part 2 show that we can replace the indicator $1_{\#\hat{\mathcal{A}}_{T-1}(H_{T-1,1})>1}$ with $1_{\#\mathcal{A}^*_{T-1}(H_{T-1,1})>1}$ and replace the sets $\tilde{\mathcal{A}}_{T-1}(H_{T-1,1})$ with $\mathcal{A}^*_{T-1}(H_{T-1,1})$ incurring a difference of the order $o_P(1)$. Finally, the argument at the end of Supplement Theorem 1.1 shows that we can replace $\mathbb{P}_n$ with P, again, up to a term of the order $o_P(1)$. Thus, the last term in (3.55) is equal to

$$\Sigma_{T-2,\infty}^{-1} P B_{T-2}^{\intercal} \Bigg(\max_{i \in \mathcal{A}^*_{T-1}(H_{T-1,1})} H_{T-1,1}^{\intercal} \left(\mathbb{V}_{T-1,n,i}(\gamma_T) + \gamma_{T-1,i} \right) - \max_{i \in \mathcal{A}^*_{T-1}(H_{T-1,1})} H_{T,1}^{\intercal} \gamma_{T-1,i} \\ - \max_{i \in \mathcal{A}^*_{T-1}(H_{T-1,1})} H_{T,1}^{\intercal} \mathbb{V}_{T-1,n,i}(\gamma_T) \Bigg) 1_{\# \mathcal{A}^*_{T-1}(H_{T-1,1}) > 1} + o_P(1).$$

The limiting distribution of $\mathbb{V}_{T-2,n}((\gamma_{T-1}, \gamma_T))$ is obtained by appeal to the continuous mapping theorem and appeal to the Cramer-Wold device. Furthermore, the limiting distribution of $\mathbb{V}_{T-2,n}((\gamma_{T-1}, \gamma_T))$ was obtained from that of $\mathbb{V}_{T-1,n}(\gamma_T)$ is a way that did not depend on the stage $t = T - 2$. That is, exactly the same arguments yield the desired limiting distribution of $\mathbb{V}_{T-3,n}((\gamma_{T-2}, \gamma_{T-1}, \gamma_T))$ from that of $\mathbb{V}_{T-2,n}((\gamma_{T-1}, \gamma_T))$. Thus, the result follows from induction. □

3.9 Appendix IV: Bias reduction for non-regular problems

In this section we briefly discuss the issue of bias reduction for non-regular problems. It is now well known that unbiased estimators do not exist for non-smooth functionals (see Robins 2004, appendix I; and Porter and Hirano 2009). Furthermore, it has been shown that attempting to reduce the bias at a non-regular point in the parameter space can dramatically inflate the variance and subsequently the MSE elsewhere in the paramter space (Doss and Sethuraman 1989; Brown and Liu 1993; Chen 2004). Here, we attempt to illustrate this phenomenon in a toy example that is relevant for medical decision making.

Suppose that X_1 and X_2 are independent normal random variables with means μ_1 and μ_2 respectively, both are assumed to have unit variance. We consider the task of estimating $\theta \triangleq \max(\mu_1, \mu_2)$ based on a single observation $X_1 = x_1$ and $X_2 = x_2$. Notice that this problem corresponds to a toy decision making problem where μ_i

denotes the mean response for patients following treatment i. The MLE is given by $\hat{\theta}_{\text{mle}} \triangleq \max(X_1, X_2)$. It is clear that the MLE suffers from upward bias since

$$\theta \triangleq \max(\mu_1, \mu_2) = \max(\mathbb{E}X_1, \mathbb{E}X_2) \leq \mathbb{E}\max(X_1, X_2).$$

It will be convenient to write the $\hat{\theta}_{\text{mle}}$ as

$$\hat{\theta}_{\text{mle}} \triangleq (X_1 + X_2)/2 + |X_1 - X_2|/2.$$

The first term on the right hand side of the above display is the UMVU estimator of θ when there is no treatment effect (e.g. $\mu_1 = \mu_2$). The second term can be seen as an estimator of the advantage of recommending treatment via the decision rule $\arg\max_{i=1,2} X_i$ compared with randomly assigning treatment according to an even odds coin flip. The thresholding estimators of Chakraborty et al. (2009) and Moodie and Richardson (2007) shrink the term $|X_1 - X_2|/2$ towards zero in an attempt to alleviate some of the bias inherent to $\hat{\theta}_{\text{mle}}$. In particular, an analogue of the soft-thresholding estimator of Chakraborty et al. (2009) for this problem is given by

$$\hat{\theta}_{\text{soft}} \triangleq (X_1 + X_2)/2 + \left[1 - \frac{\lambda}{|X_1 - X_2|}\right]_+ |X_1 - X_2|/2$$

where λ denotes a tuning parameter. An analogue of the hard-thesholding estimator of Moodie and Richardson (2007) is given by

$$\hat{\theta}_{\text{hard}} \triangleq (X_1 + X_2)/2 + 1_{|X_1 - X_2| \geq \lambda}|X_1 - X_2|/2,$$

again where λ is a tuning parameter. Notice that both estimators reduce to $\hat{\theta}_{\text{mle}}$ when $\lambda = 0$. As we will see, the bias $\hat{\theta}_{\text{mle}}$ is largest when $\mu_1 = \mu_2$. Both $\hat{\theta}_{\text{soft}}$ and $\hat{\theta}_{\text{hard}}$

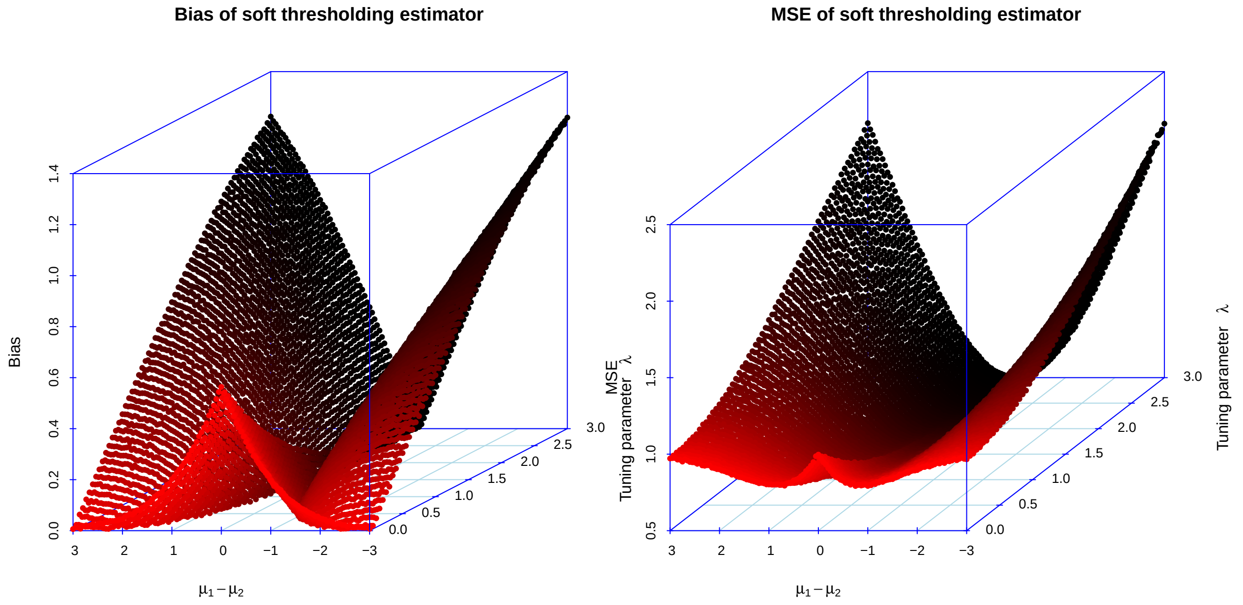

Figure 3.1: Left: The bias of $\hat{\theta}_{\text{soft}}$ as a function of effect size $\mu_1 - \mu_2$ and tuning parameter λ. Reducing the bias at $\mu_1 - \mu_2 = 0$ requires increasing λ which is seen to dramatically inflate bias elsewhere. Right: The MSE of $\hat{\theta}_{\text{soft}}$ as a function of effect size $\mu_1 - \mu_2$ and tuning parameter λ. Attempting to reduce the bias at $\mu_1 - \mu_2 = 0$ results in a modest reduction in MSE at $\mu_1 - \mu_2 = 0$ but inflates the MSE significantly elsewhere.

seek to alleviate some of this bias by shrinking $\hat{\theta}_{\text{mle}}$ towards $(X_1 + X_2)/2$ whenever $|X_1 - X_2|$ is small.

Figure (3.9) shows the bias and MSE of the soft-threshold estimator $\hat{\theta}_{\text{soft}}$ as a function of effect size $\mu_1 - \mu_2$ and tuning parameter λ. The figure shows that by increasing λ the bias at $\mu_1 - \mu_2 = 0$ decreases, however, modest increases in λ lead to dramatic increases in bias non-zero values of $\mu_1 - \mu_2$ and subsequently inflate the MSE. Figure (3.9) shows results of a similar nature for the hard-thresholding estimator $\hat{\theta}_{\text{hard}}$. These figures show that the price of bias reduction at $\mu_1 - \mu_2 = 0$ can be quite severe unless one has very strong prior knowledge about the true value of $\mu_1 - \mu_2$.

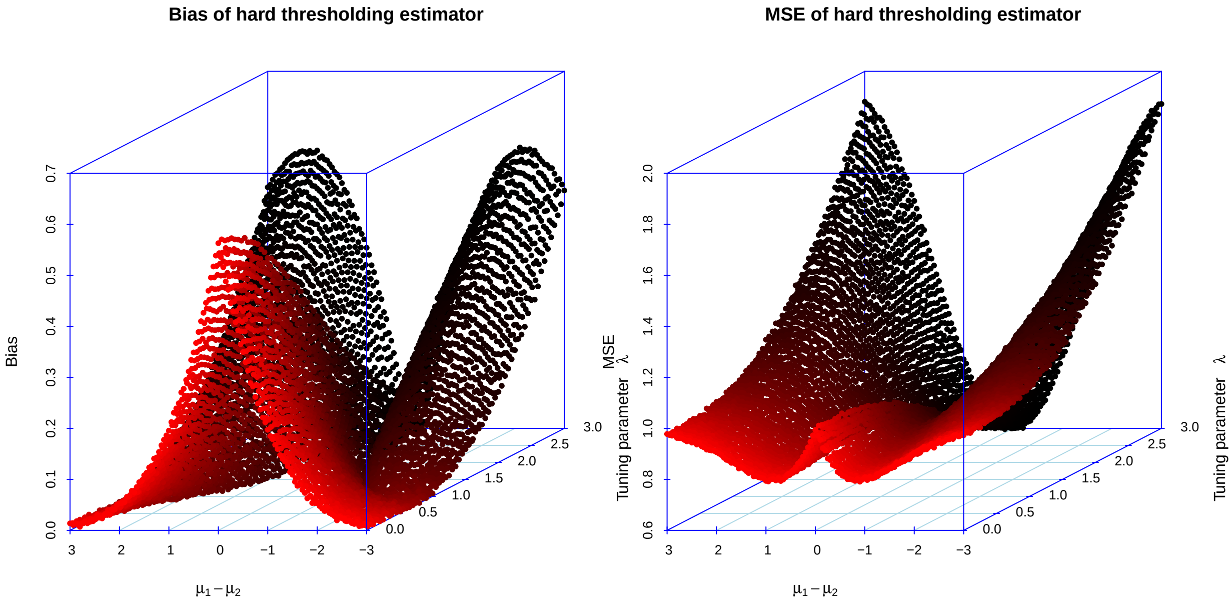

Figure 3.2: Left: The bias of $\hat{\theta}_{\text{hard}}$ as a function of effect size $\mu_1 - \mu_2$ and tuning parameter λ. Reducing the bias at $\mu_1 - \mu_2 = 0$ requires increasing λ which is seen to dramatically inflate bias elsewhere. Right: The MSE of $\hat{\theta}_{\text{hard}}$ as a function of effect size $\mu_1 - \mu_2$ and tuning parameter λ. Attempting to reduce the bias at $\mu_1 - \mu_2 = 0$ results in a modest reduction in MSE at $\mu_1 - \mu_2 = 0$ but inflates the MSE significantly elsewhere.

3.10 Appendix V: Additional empirical results

Here, we present additional empirical results for the ACI and competitors. We give results for the generative models in the main body of the paper with varying dataset sizes, for generative models with three treatments at the second stage, and for generative models with three stages of binary treatments. All of the results in this section are based on 1000 Monte Carlo repetitions, and for the ACI we use the tuning parameter $\lambda_n = \log\log n$.

3.10.1 Varying dataset size

First, we present a suite of experiments with the two-stage, two-action models presented in the main body of the paper, with varying data set size N. Tables 3.12 through 3.23 show our results.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.892*	0.908*	0.924*	0.925*	0.940	0.930*	0.936
PPE	0.926*	0.930*	0.933*	0.934*	0.934*	0.907*	0.928*
ST	0.935*	0.930*	0.889*	0.878*	0.891*	0.620*	0.687*
ACI	0.956	0.964	0.954	0.955	0.950	0.957	0.948

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.899*	0.915*	0.947	0.949	0.939	0.967	0.961
PPE	0.949	0.946	0.952	0.948	0.941	0.948	0.958
ST	0.952	0.945	0.935*	0.929*	0.935*	0.644*	0.780*
ACI	0.970	0.976	0.969	0.970	0.956	0.973	0.965

N = 500	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.892*	0.906*	0.935*	0.933*	0.929*	0.942	0.943
PPE	0.936	0.938	0.941	0.937	0.929*	0.934*	0.938
ST	0.956	0.949	0.923*	0.917*	0.910*	0.664*	0.790*
ACI	0.965	0.976	0.964	0.968	0.952	0.950	0.944

N = 1000	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.907*	0.933*	0.933*	0.943	0.944	0.945	0.951
PPE	0.949	0.938	0.949	0.947	0.952	0.942	0.949
ST	0.953	0.933*	0.944	0.934*	0.934*	0.813*	0.880*
ACI	0.968	0.980	0.968	0.971	0.961	0.946	0.951

Table 3.12: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,0}$ (intercept term) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.404*	0.404*	0.430*	0.429*	0.457	0.449*	0.450
PPE	0.376*	0.376*	0.418*	0.418*	0.451*	0.448*	0.453*
ST	0.344*	0.344*	0.427*	0.427*	0.466*	0.469*	0.474*
ACI	0.518	0.518	0.487	0.487	0.486	0.494	0.476

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.284*	0.284*	0.300	0.300	0.320	0.314	0.314
PPE	0.264	0.264	0.292	0.292	0.316	0.316	0.317
ST	0.240	0.240	0.289*	0.289*	0.319*	0.326*	0.324*
ACI	0.367	0.367	0.343	0.343	0.341	0.338	0.328

N = 500	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.218*	0.218*	0.232*	0.232*	0.248*	0.243	0.243
PPE	0.203	0.203	0.226	0.226	0.245*	0.247*	0.245
ST	0.184	0.185	0.221*	0.222*	0.245*	0.253*	0.251*
ACI	0.284	0.284	0.265	0.265	0.265	0.255	0.249

N = 1000	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.155*	0.155*	0.164*	0.164	0.175	0.171	0.171
PPE	0.144	0.144	0.159	0.160	0.173	0.173	0.172
ST	0.131	0.131*	0.156	0.156*	0.172*	0.179*	0.176*
ACI	0.202	0.202	0.188	0.188	0.187	0.174	0.172

Table 3.13: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,0}$ (intercept term) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.942	0.944	0.948	0.948	0.928*	0.942	0.939
PPE	0.946	0.946	0.945	0.945	0.931*	0.936	0.939
ST	0.946	0.946	0.950	0.950	0.941	0.941	0.941
ACI	0.964	0.966	0.958	0.957	0.941	0.947	0.940

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.942	0.947	0.952	0.950	0.948	0.946	0.958
PPE	0.944	0.946	0.953	0.953	0.943	0.942	0.956
ST	0.945	0.945	0.948	0.949	0.951	0.940	0.955
ACI	0.960	0.959	0.957	0.957	0.955	0.946	0.958

N = 500	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.948	0.951	0.954	0.953	0.948	0.952	0.953
PPE	0.948	0.950	0.955	0.953	0.951	0.951	0.952
ST	0.948	0.948	0.954	0.953	0.951	0.952	0.949
ACI	0.967	0.966	0.964	0.964	0.961	0.952	0.953

N = 1000	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.941	0.945	0.938	0.944	0.937	0.941	0.941
PPE	0.942	0.944	0.939	0.942	0.936	0.940	0.941
ST	0.945	0.947	0.944	0.943	0.941	0.939	0.945
ACI	0.963	0.961	0.955	0.955	0.945	0.941	0.941

Table 3.14: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,1}$ (main effect of history) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.331	0.331	0.333	0.333	0.379*	0.354	0.355
PPE	0.330	0.330	0.332	0.332	0.376*	0.350	0.353
ST	0.328	0.328	0.332	0.332	0.384	0.360	0.361
ACI	0.360	0.360	0.347	0.348	0.392	0.359	0.358

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.231	0.231	0.232	0.232	0.265	0.246	0.246
PPE	0.230	0.230	0.231	0.231	0.263	0.245	0.246
ST	0.229	0.229	0.231	0.231	0.266	0.250	0.249
ACI	0.251	0.251	0.242	0.242	0.275	0.248	0.247

N = 500	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.178	0.178	0.179	0.179	0.205	0.190	0.190
PPE	0.178	0.178	0.178	0.178	0.204	0.190	0.190
ST	0.177	0.177	0.178	0.178	0.205	0.193	0.192
ACI	0.194	0.194	0.187	0.187	0.213	0.191	0.191

N = 1000	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.126	0.126	0.126	0.126	0.145	0.134	0.134
PPE	0.125	0.125	0.126	0.126	0.144	0.134	0.134
ST	0.124	0.124	0.125	0.125	0.144	0.135	0.135
ACI	0.137	0.137	0.132	0.132	0.150	0.134	0.134

Table 3.15: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,1}$ (main effect of history) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.944	0.943	0.948	0.949	0.944	0.952	0.956
PPE	0.944	0.945	0.951	0.951	0.941	0.947	0.954
ST	0.946	0.946	0.950	0.949	0.955	0.950	0.952
ACI	0.963	0.963	0.959	0.959	0.955	0.953	0.957

N = 300	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.960	0.958	0.953	0.954	0.949	0.950	0.948
PPE	0.957	0.956	0.954	0.954	0.945	0.945	0.948
ST	0.954	0.954	0.952	0.951	0.943	0.946	0.949
ACI	0.975	0.975	0.964	0.963	0.955	0.951	0.949

N = 500	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.953	0.952	0.945	0.947	0.936	0.951	0.941
PPE	0.953	0.954	0.944	0.945	0.938	0.951	0.941
ST	0.947	0.947	0.945	0.945	0.938	0.945	0.938
ACI	0.966	0.966	0.956	0.956	0.948	0.952	0.941

N = 1000	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.954	0.955	0.951	0.955	0.956	0.957	0.953
PPE	0.954	0.955	0.952	0.954	0.959	0.957	0.953
ST	0.953	0.953	0.951	0.952	0.954	0.959	0.954
ACI	0.967	0.969	0.959	0.961	0.965	0.958	0.953

Table 3.16: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}$ (interaction between history and treatment) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.331	0.331	0.333	0.332	0.363	0.354	0.355
PPE	0.330	0.330	0.332	0.332	0.361	0.350	0.353
ST	0.328	0.328	0.332	0.332	0.366	0.359	0.360
ACI	0.360	0.360	0.347	0.347	0.378	0.359	0.358

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.231	0.231	0.231	0.231	0.254	0.246	0.246
PPE	0.230	0.230	0.231	0.231	0.252	0.244	0.246
ST	0.228	0.228	0.230	0.230	0.254	0.250	0.249
ACI	0.251	0.250	0.241	0.241	0.264	0.248	0.247

N = 500	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.178	0.178	0.178	0.178	0.197	0.190	0.190
PPE	0.177	0.177	0.178	0.178	0.196	0.189	0.190
ST	0.176	0.176	0.178	0.178	0.196	0.193	0.191
ACI	0.194	0.194	0.186	0.186	0.205	0.191	0.191

N = 1000	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.126	0.126	0.126	0.126	0.139	0.134	0.134
PPE	0.125	0.125	0.126	0.126	0.138	0.134	0.134
ST	0.125	0.125	0.125	0.125	0.139	0.135	0.134
ACI	0.137	0.137	0.132	0.132	0.145	0.134	0.134

Table 3.17: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}$ (interaction between history and treatment) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.934*	0.935*	0.930*	0.933*	0.938	0.928*	0.939
PPE	0.931*	0.940	0.938	0.940	0.946	0.912*	0.931*
ST	0.948	0.945	0.938	0.942	0.952	0.943	0.919*
ACI	0.992	0.992	0.968	0.972	0.957	0.955	0.950

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.952	0.952	0.948	0.952	0.943	0.936	0.941
PPE	0.951	0.952	0.960	0.959	0.956	0.907*	0.944
ST	0.951	0.949	0.938	0.941	0.949	0.951	0.920*
ACI	0.994	0.994	0.975	0.976	0.962	0.957	0.950

N = 500	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.947	0.944	0.947	0.947	0.943	0.946	0.944
PPE	0.952	0.945	0.950	0.951	0.940	0.919*	0.945
ST	0.965	0.965	0.953	0.959	0.951	0.927*	0.910*
ACI	0.992	0.992	0.976	0.980	0.956	0.958	0.947

N = 1000	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.948	0.949	0.934*	0.939	0.950	0.954	0.951
PPE	0.948	0.949	0.948	0.945	0.952	0.941	0.948
ST	0.956	0.955	0.959	0.955	0.954	0.935*	0.924*
ACI	0.998	0.995	0.972	0.973	0.963	0.954	0.951

Table 3.18: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,1}$ (main effect of treatment) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.385*	0.385*	0.430*	0.430*	0.457	0.436*	0.451
PPE	0.365*	0.366	0.419	0.419	0.452	0.418*	0.452*
ST	0.339	0.339	0.426	0.427	0.469	0.436	0.480*
ACI	0.502	0.502	0.488	0.488	0.487	0.475	0.477

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.269	0.269	0.300	0.300	0.320	0.309	0.313
PPE	0.256	0.256	0.292	0.292	0.316	0.297*	0.317
ST	0.237	0.237	0.289	0.289	0.320	0.313	0.327*
ACI	0.354	0.354	0.342	0.342	0.341	0.327	0.327

N = 500	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.208	0.208	0.232	0.232	0.248	0.242	0.244
PPE	0.197	0.197	0.226	0.226	0.245	0.234*	0.245
ST	0.182	0.183	0.222	0.222	0.246	0.252*	0.253*
ACI	0.275	0.275	0.265	0.265	0.265	0.250	0.250

N = 1000	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.147	0.147	0.164*	0.164	0.175	0.171	0.171
PPE	0.139	0.139	0.160	0.160	0.173	0.170	0.172
ST	0.129	0.129	0.156	0.156	0.172	0.184*	0.177*
ACI	0.195	0.195	0.188	0.188	0.187	0.173	0.173

Table 3.19: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,1}$ (main effect of treatment) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.942	0.941	0.940	0.942	0.943	0.929*	0.941
PPE	0.941	0.938	0.945	0.943	0.937	0.917*	0.935*
ST	0.943	0.943	0.932*	0.934*	0.940	0.934*	0.928*
ACI	0.984	0.985	0.963	0.964	0.948	0.946	0.949

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.953	0.953	0.954	0.959	0.955	0.953	0.961
PPE	0.958	0.955	0.965	0.966	0.956	0.937	0.959
ST	0.956	0.958	0.945	0.950	0.958	0.956	0.934*
ACI	0.988	0.989	0.977	0.979	0.961	0.966	0.966

N = 500	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.951	0.951	0.948	0.950	0.952	0.954	0.954
PPE	0.952	0.951	0.953	0.952	0.953	0.938	0.953
ST	0.950	0.950	0.956	0.957	0.950	0.938	0.931*
ACI	0.988	0.988	0.968	0.973	0.958	0.957	0.954

N = 1000	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.954	0.949	0.945	0.946	0.953	0.948	0.949
PPE	0.958	0.958	0.952	0.951	0.955	0.942	0.948
ST	0.966	0.965	0.955	0.960	0.956	0.938	0.934*
ACI	0.992	0.991	0.972	0.978	0.962	0.951	0.950

Table 3.20: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0} + \beta^*_{1,1,1}$ (effect of action for history = 1) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.506	0.506	0.542	0.542	0.580	0.559*	0.571
PPE	0.491	0.491	0.533	0.533	0.578	0.544*	0.571*
ST	0.471	0.471	0.539*	0.539*	0.600	0.563*	0.598*
ACI	0.622	0.622	0.600	0.600	0.596	0.598	0.595

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.355	0.355	0.378	0.378	0.406	0.394	0.397
PPE	0.344	0.344	0.372	0.372	0.404	0.383	0.400
ST	0.329	0.329	0.369	0.369	0.412	0.399	0.410*
ACI	0.439	0.439	0.421	0.421	0.417	0.411	0.409

N = 500	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.273	0.274	0.293	0.293	0.315	0.307	0.308
PPE	0.265	0.265	0.288	0.288	0.314	0.301	0.310
ST	0.254	0.254	0.284	0.284	0.318	0.317	0.315*
ACI	0.340	0.340	0.327	0.327	0.324	0.315	0.314

N = 1000	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.193	0.194	0.207	0.207	0.222	0.217	0.217
PPE	0.187	0.187	0.203	0.203	0.222	0.216	0.217
ST	0.180	0.180	0.200	0.200	0.223	0.228	0.221*
ACI	0.241	0.242	0.231	0.231	0.229	0.218	0.218

Table 3.21: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0} + \beta^*_{1,1,1}$ (effect of action for history = 1) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.958	0.954	0.944	0.945	0.935*	0.943	0.949
PPE	0.954	0.952	0.950	0.950	0.938	0.935*	0.941
ST	0.964	0.964	0.940	0.943	0.938	0.951	0.929*
ACI	0.985	0.985	0.970	0.972	0.960	0.960	0.953

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.949	0.951	0.941	0.942	0.947	0.936	0.945
PPE	0.949	0.952	0.942	0.943	0.946	0.920*	0.945
ST	0.950	0.950	0.937	0.942	0.943	0.940	0.923*
ACI	0.984	0.985	0.964	0.966	0.965	0.951	0.949

N = 500	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.943	0.944	0.934*	0.942	0.933*	0.944	0.944
PPE	0.949	0.947	0.947	0.950	0.942	0.927*	0.946
ST	0.963	0.961	0.943	0.946	0.937	0.929*	0.920*
ACI	0.986	0.985	0.962	0.967	0.953	0.950	0.948

N = 1000	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.947	0.940	0.937	0.944	0.944	0.950	0.946
PPE	0.947	0.946	0.955	0.951	0.944	0.948	0.946
ST	0.948	0.946	0.961	0.961	0.953	0.936	0.929*
ACI	0.989	0.990	0.971	0.974	0.962	0.953	0.946

Table 3.22: Monte Carlo estimates of coverage probabilities of confidence intervals for $-\beta^*_{1,1,0}+\beta^*_{1,1,1}$ (effect of action for history = -1) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.508	0.508	0.544	0.544	0.587*	0.563	0.575
PPE	0.491	0.491	0.534	0.534	0.577	0.545*	0.573
ST	0.471	0.471	0.541	0.542	0.588	0.566	0.601*
ACI	0.624	0.624	0.601	0.601	0.630	0.600	0.600

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.354	0.354	0.379	0.378	0.410	0.398	0.400
PPE	0.343	0.343	0.372	0.372	0.404	0.385*	0.402
ST	0.329	0.329	0.369	0.370	0.404	0.403	0.412*
ACI	0.439	0.439	0.420	0.420	0.442	0.415	0.413

N = 500	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.273	0.273	0.293*	0.293	0.318*	0.309	0.310
PPE	0.265	0.265	0.288	0.288	0.313	0.302*	0.311
ST	0.253	0.253	0.284	0.284	0.311	0.318*	0.318*
ACI	0.340	0.340	0.325	0.325	0.343	0.316	0.315

N = 1000	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.193	0.193	0.206	0.206	0.224	0.218	0.218
PPE	0.187	0.187	0.203	0.203	0.221	0.217	0.218
ST	0.179	0.179	0.200	0.200	0.219	0.229	0.223*
ACI	0.241	0.241	0.230	0.230	0.242	0.219	0.219

Table 3.23: Monte Carlo estimates of coverage probabilities of confidence intervals for $-\beta^*_{1,1,0}+\beta^*_{1,1,1}$ (effect of action for history = -1) at the 95% nominal level. Generative models have two stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300, 500, and 1000 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

3.10.2 Models with ternary actions

Here, we present results using a suite of examples similar to those of Chakraborty et al. (2009), but that have three possible treatments at the second stage. These models are defined as follows:

- $X_i \in \{-1, 1\}$ for $i \in \{1, 2\}$, $A_1 \in \{-1, 1\}$, and $A_2 \in \{(0, -0.5)^\intercal, (-1, 0.5)^\intercal, (1, 0.5)^\intercal\}$
- $P(A_1 = 1) = P(A_1 = -1) = 1/2$,
  $P(A_2 = (0, -1)^\intercal) = P(A_2 = (-1, 0.5)^\intercal) = P(A_2 = (1, 0.5)^\intercal) = 1/3$
- $P(X_1 = 1) = P(X_1 = -1) = 1/2$, $P(X_2 = 1|X_1, A_1) = \text{expit}(\delta_1 X_1 + \delta_2 A_1)$
- $Y_1 \triangleq 0$,
  $Y_2 = \gamma_1 + \gamma_2 X_1 + \gamma_3 A_1 + \gamma_4 X_1 A_1 + (\gamma_5, \gamma_6) A_2 + X_2(\gamma_7, \gamma_8) A_2 + A_1(\gamma_9, \gamma_{10}) A_2 + \epsilon$,
  $\epsilon \sim N(0, 1)$

where $\text{expit}(x) = \mathrm{e}^x/(1+\mathrm{e}^x)$. This class is parameterized by twelve values $\gamma_1, \gamma_2, ..., \gamma_{10}, \delta_1, \delta_2$. The analysis model uses histories defined by:

$$\begin{aligned} H_{2,0} &= (1, X_1, A_1, X_1 A_1, X_2)^\intercal && (3.56) \\ H_{2,1} &= (1, X_2, A_1)^\intercal && (3.57) \\ H_{1,0} &= (1, X_1)^\intercal && (3.58) \\ H_{1,1} &= (1, X_1)^\intercal. && (3.59) \end{aligned}$$

Our working models are given by $Q_2(H_2, A_2; \beta_2) \triangleq H_{2,0}^\intercal \beta_{2,0} + H_{2,1}^\intercal \beta_{2,1,1} A_{2,1} + H_{2,1}^\intercal \beta_{2,1,2} A_{2,2}$ and $Q_1(H_1, A_1; \beta_1) \triangleq H_{1,0}^\intercal \beta_{1,0} + H_{1,1}^\intercal \beta_{1,1} A_1$. In Table 3.10.2, for each of these models we give the probability p of generating a history where each of the three possible treatments at the second stage have exactly the same effect. This is analogous to

Example	γ	δ	Regularity
1	$(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)^{\intercal}$	$(0.5, 0.5)^{\intercal}$	$p = 1, \phi = 0/0$
2	$(0, 0, 0, 0, 0.01, 0.01, 0, 0, 0, 0)^{\intercal}$	$(0.5, 0.5)^{\intercal}$	$p = 0, \phi = \infty$
3	$(0, 0, -0.5, 0, 0.5, 0.5, 0, 0, 0.5, 0.5)^{\intercal}$	$(0.5, 0.5)^{\intercal}$	$p = 1/2, \phi = 1.0$
4	$(0, 0, -0.5, 0, 0.5, 0.5, 0, 0, 0.49, 0.49)^{\intercal}$	$(0.5, 0.5)^{\intercal}$	$p = 0, \phi = 1.0204$
5	$(0, 0, -0.5, 0, 1.00, 1.00, 0.5, 0.5, 0.5, 0.5)^{\intercal}$	$(1.0, 0.0)^{\intercal}$	$p = 1/4, \phi = 1.4142$
6	$(0, 0, -0.5, 0, 0.25, 0.25, 0.5, 0.5, 0.5, 0.5)^{\intercal}$	$(0.1, 0.1)^{\intercal}$	$p = 0, \phi = 0.3451$
7	$(0, 0, -0.25, 0, 0.75, 0.75, 0.5, 0.5, 0.5, 0.5)^{\intercal}$	$(0.1, 0.1)^{\intercal}$	$p = 0, \phi = 1.035$

Table 3.24: Parameters indexing the example models.

having the second stage action show no effect in a binary model. Furthermore, because of the Helmert encoding we have used in our analysis models, and because of the structure of γ, it happens that the standardized effect size of treatment 1 versus treatment 2, treatment 1 versus treatment 3, and treatment 2 versus treatment 3 are all exactly equal in our examples. We report this as ϕ in Table 3.10.2. Tables 3.25 through 3.36 detail our results.

N = 150	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.836*	0.868*	0.929*	0.930*	0.928*	0.926*	0.930*
PPE	0.860*	0.881*	0.926*	0.926*	0.923*	0.919*	0.929*
ST	0.938	0.946	0.853*	0.854*	0.876*	0.591*	0.801*
ACI	0.909*	0.931*	0.949	0.949	0.953	0.950	0.951

N = 300	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.833*	0.879*	0.953	0.953	0.954	0.937	0.931*
PPE	0.848*	0.898*	0.948	0.949	0.947	0.937	0.926*
ST	0.930*	0.945	0.896*	0.896*	0.913*	0.654*	0.861*
ACI	0.898*	0.934*	0.960	0.961	0.961	0.948	0.944

Table 3.25: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,0}$ (intercept term) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.503*	0.503*	0.655*	0.655*	0.719*	0.598*	0.685*
PPE	0.499*	0.499*	0.649*	0.650*	0.710*	0.598*	0.675*
ST	0.438	0.438	0.717*	0.719*	0.769*	0.608*	0.751*
ACI	0.579*	0.579*	0.720	0.720	0.782	0.671	0.752

N = 300	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.345*	0.346*	0.458	0.458	0.502	0.420	0.481*
PPE	0.343*	0.343*	0.457	0.458	0.500	0.419	0.473*
ST	0.298*	0.298	0.477*	0.478*	0.520*	0.432*	0.510*
ACI	0.383*	0.383*	0.490	0.490	0.531	0.450	0.515

Table 3.26: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,0}$ (intercept term) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.952	0.952	0.955	0.956	0.941	0.963	0.969
PPE	0.953	0.952	0.954	0.955	0.944	0.961	0.966
ST	0.952	0.952	0.954	0.954	0.950	0.962	0.968
ACI	0.987	0.988	0.979	0.977	0.965	0.978	0.982

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.963	0.964	0.961	0.961	0.957	0.954	0.951
PPE	0.963	0.964	0.961	0.961	0.958	0.954	0.951
ST	0.967	0.965	0.960	0.960	0.959	0.953	0.950
ACI	0.977	0.977	0.971	0.971	0.968	0.965	0.964

Table 3.27: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,1}$ (main effect of history) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.339	0.340	0.346	0.346	0.454	0.384	0.398
PPE	0.339	0.339	0.346	0.346	0.452	0.382	0.397
ST	0.337	0.337	0.347	0.347	0.466	0.391	0.406
ACI	0.398	0.398	0.386	0.386	0.492	0.423	0.434

N = 300	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.234	0.234	0.238	0.238	0.317	0.264	0.275
PPE	0.234	0.234	0.238	0.238	0.317	0.264	0.274
ST	0.233	0.233	0.238	0.238	0.322	0.270	0.279
ACI	0.263	0.262	0.257	0.257	0.334	0.281	0.291

Table 3.28: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,1}$ (main effect of history) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.942	0.941	0.950	0.950	0.947	0.941	0.952
PPE	0.944	0.945	0.950	0.950	0.949	0.938	0.950
ST	0.948	0.948	0.949	0.949	0.952	0.942	0.950
ACI	0.969	0.969	0.962	0.962	0.966	0.960	0.966

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.948	0.949	0.953	0.953	0.962	0.946	0.961
PPE	0.949	0.949	0.953	0.953	0.961	0.946	0.959
ST	0.950	0.951	0.952	0.951	0.962	0.944	0.963
ACI	0.980	0.980	0.968	0.967	0.970	0.955	0.972

Table 3.29: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}$ (interaction between history and treatment) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.339	0.339	0.338	0.338	0.419	0.385	0.399
PPE	0.338	0.339	0.338	0.338	0.422	0.383	0.398
ST	0.338	0.338	0.338	0.338	0.426	0.392	0.406
ACI	0.397	0.397	0.377	0.377	0.456	0.424	0.434

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.232	0.232	0.231	0.231	0.290	0.263	0.274
PPE	0.232	0.232	0.231	0.231	0.291	0.263	0.273
ST	0.232	0.232	0.230	0.230	0.293	0.269	0.278
ACI	0.261	0.261	0.249	0.249	0.306	0.280	0.290

Table 3.30: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}$ (interaction between history and treatment) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.949	0.952	0.937	0.938	0.942	0.930*	0.936
PPE	0.948	0.949	0.935*	0.936	0.938	0.915*	0.933*
ST	0.945	0.946	0.939	0.939	0.938	0.915*	0.933*
ACI	0.982	0.983	0.960	0.960	0.964	0.953	0.959

N = 300	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.957	0.953	0.950	0.950	0.954	0.938	0.946
PPE	0.957	0.955	0.950	0.950	0.954	0.920*	0.945
ST	0.953	0.951	0.950	0.950	0.954	0.932*	0.949
ACI	0.975	0.972	0.966	0.966	0.965	0.953	0.966

Table 3.31: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,1}$ (main effect of treatment) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.406	0.406	0.468	0.468	0.551	0.506*	0.532
PPE	0.402	0.402	0.469*	0.469	0.550	0.499*	0.528*
ST	0.400	0.400	0.473	0.473	0.566	0.530*	0.552*
ACI	0.475	0.475	0.521	0.521	0.602	0.569	0.589

N = 300	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.279	0.279	0.326	0.326	0.385	0.358	0.373
PPE	0.277	0.277	0.326	0.326	0.385	0.354*	0.369
ST	0.274	0.275	0.326	0.326	0.392	0.375*	0.383
ACI	0.314	0.314	0.350	0.350	0.408	0.384	0.402

Table 3.32: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,1}$ (main effect of treatment) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.942	0.940	0.941	0.941	0.945	0.939	0.941
PPE	0.944	0.943	0.939	0.938	0.944	0.927*	0.940
ST	0.947	0.948	0.940	0.941	0.946	0.924*	0.939
ACI	0.976	0.977	0.961	0.962	0.964	0.957	0.963

N = 300	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.953	0.956	0.951	0.951	0.952	0.952	0.944
PPE	0.953	0.956	0.951	0.951	0.951	0.947	0.941
ST	0.954	0.957	0.954	0.954	0.950	0.943	0.941
ACI	0.976	0.974	0.965	0.965	0.960	0.965	0.954

Table 3.33: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0} + \beta^*_{1,1,1}$ (effect of action for history = 1) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.529	0.529	0.577	0.577	0.693	0.631	0.664
PPE	0.525	0.526	0.578	0.578	0.689	0.626*	0.660
ST	0.522	0.523	0.582	0.582	0.709	0.655*	0.686
ACI	0.609	0.609	0.640	0.640	0.745	0.703	0.729

N = 300	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.363	0.363	0.398	0.398	0.483	0.441	0.462
PPE	0.361	0.362	0.398	0.398	0.482	0.438	0.459
ST	0.360	0.360	0.399	0.399	0.489	0.458	0.472
ACI	0.403	0.403	0.426	0.426	0.508	0.472	0.495

Table 3.34: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0} + \beta^*_{1,1,1}$ (effect of action for history = 1) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.940	0.939	0.930*	0.931*	0.947	0.937	0.939
PPE	0.938	0.939	0.930*	0.929*	0.946	0.930*	0.934*
ST	0.934*	0.934*	0.933*	0.933*	0.947	0.930*	0.937
ACI	0.965	0.963	0.960	0.960	0.965	0.956	0.964

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.943	0.945	0.948	0.948	0.958	0.938	0.957
PPE	0.941	0.943	0.948	0.948	0.960	0.927*	0.954
ST	0.945	0.947	0.951	0.951	0.960	0.933*	0.954
ACI	0.968	0.964	0.966	0.966	0.970	0.955	0.970

Table 3.35: Monte Carlo estimates of coverage probabilities of confidence intervals for $-\beta^*_{1,1,0} + \beta^*_{1,1,1}$ (effect of action for history = -1) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.530	0.530	0.575*	0.575*	0.690	0.639	0.664
PPE	0.526	0.526	0.577*	0.577*	0.695	0.632*	0.660*
ST	0.524*	0.524*	0.579*	0.580*	0.706	0.664*	0.684
ACI	0.610	0.609	0.638	0.638	0.758	0.710	0.730

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.363	0.363	0.399	0.399	0.482	0.447	0.463
PPE	0.361	0.361	0.400	0.400	0.483	0.443*	0.460
ST	0.359	0.359	0.400	0.400	0.489	0.464*	0.473
ACI	0.402	0.403	0.428	0.428	0.511	0.477	0.497

Table 3.36: Monte Carlo estimates of coverage probabilities of confidence intervals for $-\beta^*_{1,1,0} + \beta^*_{1,1,1}$ (effect of action for history = -1) at the 95% nominal level. Generative models two stages and **three** actions at the second stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

3.10.3 Models with three stages

Here, we present results using another suite of examples, again similar to those of Chakraborty et al. (2009), but that have three stages of treatment, with binary treatments at each stage. These models are defined as follows:

- $X_i \in \{-1, 1\}$, $A_i \in \{-1, 1\}$ for $i \in \{1, 2, 3\}$
- $P(A_i = 1) = P(A_i = -1) = 0.5$ for $i \in \{1, 2, 3\}$
- $P(X_1 = 1) = P(X_1 = -1) = 1/2$,

  $P(X_{i+1} = 1|X_i, A_i) = \text{expit}(\delta_1 X_i + \delta_2 A_i)$ for $i \in \{1, 2\}$
- $Y_1 \triangleq Y_2 \triangleq 0$

  $$\begin{aligned} Y_3 = {} & \gamma_1 + \gamma_2 X_1 + \gamma_3 A_1 + \gamma_4 X_1 A_1 + \\ & \gamma_5 A_2 + \gamma_6 X_2 A_2 + \gamma_7 A_1 A_2 + \\ & \gamma_5 A_3 + \gamma_6 X_3 A_3 + \gamma_7 A_2 A_3 + \epsilon \end{aligned}$$

  $\epsilon \sim N(0, 1)$

where $\text{expit}(x) = \text{e}^x/(1+\text{e}^x)$. This class is parameterized by nine values $\gamma_1, \gamma_2, ..., \gamma_7, \delta_1, \delta_2$. The analysis model uses histories defined by:

$$H_{3,0} = (1, X_1, A_1, X_1A_1, X_2, A_2, X_2A_2, A_1A_2, X_3)^\intercal \tag{3.60}$$

$$H_{3,1} = (1, X_3, A_2)^\intercal \tag{3.61}$$

$$H_{2,0} = (1, X_1, A_1, X_1A_1, X_2)^\intercal \tag{3.62}$$

$$H_{2,1} = (1, X_2, A_1)^\intercal \tag{3.63}$$

$$H_{1,0} = (1, X_1)^\intercal \tag{3.64}$$

$$H_{1,1} = (1, X_1)^\intercal. \tag{3.65}$$

Example	True Effect of A_2 on Stage 2 Value	Stage 2 Regularity	Stage 3 Regularity
1	$A_2(0 + 0X_2 + 0A_1)$	$p = 1, \phi = 0/0$	$p = 1, \phi = 0/0$
2	$A_2(0.01 + 0X_2 + 0A_1)$	$p = 0, \phi = \infty$	$p = 0, \phi = \infty$
3	$A_2(1.0 + 0.5X_2 + 0A_1)$	$p = 0, \phi = 2.01$	$p = 1/2, \phi = 1.003$
4	$A_2(0.99 + 0.49X_2 + 0A_1)$	$p = 0, \phi = 2.03$	$p = 0, \phi = 1.014$
5	$A_2(1.5 + 0.5X_2 + 0.5A_1)$	$p = 0, \phi = 1.92$	$p = 1/4, \phi = 1.40$
6	$A_2(0.381 + 0.500X_2 + 0.519A_1)$	$p = 0, \phi = 0.48$	$p = 0, \phi = 0.349$
7	$A_2(1.144 + 0.500X_2 + 0.506A_1)$	$p = 0, \phi = 1.46$	$p = 0, \phi = 1.05$

The values of the constants $\gamma_1, ..., \gamma_7$ and δ_1, δ_2 in Examples 1 through 7 are the same as those used in the corresponding two-stage binary action models. Since the third stage of these models has the same structure as the second stage of the models in Chakraborty et al. (2009), the non-regularity properties of the final stages in both suites of examples share the same non-regularity properties.

At stage 2 in these models, the true effect of A_2 on the stage 2 value—assuming we choose the optimal treatment at Stage 3—is given by in Table 3.10.3, along with the regularity information for the final stage. Note that in these models, the probability p of having no effect of A_2 is 1 for Example 1, and 0 for all the other examples.

Tables 3.37 through 3.48 detail our results for these models.

N = 150	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.836*	0.868*	0.929*	0.930*	0.928*	0.926*	0.930*
PPE	0.860*	0.881*	0.926*	0.926*	0.923*	0.919*	0.929*
ST	0.938	0.946	0.853*	0.854*	0.876*	0.591*	0.801*
ACI	0.909*	0.931*	0.949	0.949	0.953	0.950	0.951

N = 300	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.833*	0.879*	0.953	0.953	0.954	0.937	0.931*
PPE	0.848*	0.898*	0.948	0.949	0.947	0.937	0.926*
ST	0.930*	0.945	0.896*	0.896*	0.913*	0.654*	0.861*
ACI	0.898*	0.934*	0.960	0.961	0.961	0.948	0.944

Table 3.37: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,0}$ (intercept term) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.503*	0.503*	0.655*	0.655*	0.719*	0.598*	0.685*
PPE	0.499*	0.499*	0.649*	0.650*	0.710*	0.598*	0.675*
ST	0.438	0.438	0.717*	0.719*	0.769*	0.608*	0.751*
ACI	0.579*	0.579*	0.720	0.720	0.782	0.671	0.752

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.345*	0.346*	0.458	0.458	0.502	0.420	0.481*
PPE	0.343*	0.343*	0.457	0.458	0.500	0.419	0.473*
ST	0.298*	0.298	0.477*	0.478*	0.520*	0.432*	0.510*
ACI	0.383*	0.383*	0.490	0.490	0.531	0.450	0.515

Table 3.38: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,0}$ (intercept term) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.952	0.952	0.955	0.956	0.941	0.963	0.969
PPE	0.953	0.952	0.954	0.955	0.944	0.961	0.966
ST	0.952	0.952	0.954	0.954	0.950	0.962	0.968
ACI	0.987	0.988	0.979	0.977	0.965	0.978	0.982

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.963	0.964	0.961	0.961	0.957	0.954	0.951
PPE	0.963	0.964	0.961	0.961	0.958	0.954	0.951
ST	0.967	0.965	0.960	0.960	0.959	0.953	0.950
ACI	0.977	0.977	0.971	0.971	0.968	0.965	0.964

Table 3.39: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,1}$ (main effect of history) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.339	0.340	0.346	0.346	0.454	0.384	0.398
PPE	0.339	0.339	0.346	0.346	0.452	0.382	0.397
ST	0.337	0.337	0.347	0.347	0.466	0.391	0.406
ACI	0.398	0.398	0.386	0.386	0.492	0.423	0.434

N = 300	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.234	0.234	0.238	0.238	0.317	0.264	0.275
PPE	0.234	0.234	0.238	0.238	0.317	0.264	0.274
ST	0.233	0.233	0.238	0.238	0.322	0.270	0.279
ACI	0.263	0.262	0.257	0.257	0.334	0.281	0.291

Table 3.40: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,0,1}$ (main effect of history) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.942	0.941	0.950	0.950	0.947	0.941	0.952
PPE	0.944	0.945	0.950	0.950	0.949	0.938	0.950
ST	0.948	0.948	0.949	0.949	0.952	0.942	0.950
ACI	0.969	0.969	0.962	0.962	0.966	0.960	0.966

N = 300	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.948	0.949	0.953	0.953	0.962	0.946	0.961
PPE	0.949	0.949	0.953	0.953	0.961	0.946	0.959
ST	0.950	0.951	0.952	0.951	0.962	0.944	0.963
ACI	0.980	0.980	0.968	0.967	0.970	0.955	0.972

Table 3.41: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}$ (interaction between history and treatment) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.339	0.339	0.338	0.338	0.419	0.385	0.399
PPE	0.338	0.339	0.338	0.338	0.422	0.383	0.398
ST	0.338	0.338	0.338	0.338	0.426	0.392	0.406
ACI	0.397	0.397	0.377	0.377	0.456	0.424	0.434

N = 300	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.232	0.232	0.231	0.231	0.290	0.263	0.274
PPE	0.232	0.232	0.231	0.231	0.291	0.263	0.273
ST	0.232	0.232	0.230	0.230	0.293	0.269	0.278
ACI	0.261	0.261	0.249	0.249	0.306	0.280	0.290

Table 3.42: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0}$ (interaction between history and treatment) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.949	0.952	0.937	0.938	0.942	0.930*	0.936
PPE	0.948	0.949	0.935*	0.936	0.938	0.915*	0.933*
ST	0.945	0.946	0.939	0.939	0.938	0.915*	0.933*
ACI	0.982	0.983	0.960	0.960	0.964	0.953	0.959

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.957	0.953	0.950	0.950	0.954	0.938	0.946
PPE	0.957	0.955	0.950	0.950	0.954	0.920*	0.945
ST	0.953	0.951	0.950	0.950	0.954	0.932*	0.949
ACI	0.975	0.972	0.966	0.966	0.965	0.953	0.966

Table 3.43: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,1}$ (main effect of treatment) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.406	0.406	0.468	0.468	0.551	0.506*	0.532
PPE	0.402	0.402	0.469*	0.469	0.550	0.499*	0.528*
ST	0.400	0.400	0.473	0.473	0.566	0.530*	0.552*
ACI	0.475	0.475	0.521	0.521	0.602	0.569	0.589

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.279	0.279	0.326	0.326	0.385	0.358	0.373
PPE	0.277	0.277	0.326	0.326	0.385	0.354*	0.369
ST	0.274	0.275	0.326	0.326	0.392	0.375*	0.383
ACI	0.314	0.314	0.350	0.350	0.408	0.384	0.402

Table 3.44: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,1}$ (main effect of treatment) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.942	0.940	0.941	0.941	0.945	0.939	0.941
PPE	0.944	0.943	0.939	0.938	0.944	0.927*	0.940
ST	0.947	0.948	0.940	0.941	0.946	0.924*	0.939
ACI	0.976	0.977	0.961	0.962	0.964	0.957	0.963

N = 300	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.953	0.956	0.951	0.951	0.952	0.952	0.944
PPE	0.953	0.956	0.951	0.951	0.951	0.947	0.941
ST	0.954	0.957	0.954	0.954	0.950	0.943	0.941
ACI	0.976	0.974	0.965	0.965	0.960	0.965	0.954

Table 3.45: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0} + \beta^*_{1,1,1}$ (effect of action for history = 1) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.529	0.529	0.577	0.577	0.693	0.631	0.664
PPE	0.525	0.526	0.578	0.578	0.689	0.626*	0.660
ST	0.522	0.523	0.582	0.582	0.709	0.655*	0.686
ACI	0.609	0.609	0.640	0.640	0.745	0.703	0.729

N = 300	Ex. 1 NR	Ex. 2 NNR	Ex. 3 NR	Ex. 4 NNR	Ex. 5 NR	Ex. 6 R	Ex. 7 R
CPB	0.363	0.363	0.398	0.398	0.483	0.441	0.462
PPE	0.361	0.362	0.398	0.398	0.482	0.438	0.459
ST	0.360	0.360	0.399	0.399	0.489	0.458	0.472
ACI	0.403	0.403	0.426	0.426	0.508	0.472	0.495

Table 3.46: Monte Carlo estimates of coverage probabilities of confidence intervals for $\beta^*_{1,1,0} + \beta^*_{1,1,1}$ (effect of action for history = 1) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.940	0.939	0.930*	0.931*	0.947	0.937	0.939
PPE	0.938	0.939	0.930*	0.929*	0.946	0.930*	0.934*
ST	0.934*	0.934*	0.933*	0.933*	0.947	0.930*	0.937
ACI	0.965	0.963	0.960	0.960	0.965	0.956	0.964

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.943	0.945	0.948	0.948	0.958	0.938	0.957
PPE	0.941	0.943	0.948	0.948	0.960	0.927*	0.954
ST	0.945	0.947	0.951	0.951	0.960	0.933*	0.954
ACI	0.968	0.964	0.966	0.966	0.970	0.955	0.970

Table 3.47: Monte Carlo estimates of coverage probabilities of confidence intervals for $-\beta^*_{1,1,0} + \beta^*_{1,1,1}$ (effect of action for history = -1) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

N = 150	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.530	0.530	0.575*	0.575*	0.690	0.639	0.664
PPE	0.526	0.526	0.577*	0.577*	0.695	0.632*	0.660*
ST	0.524*	0.524*	0.579*	0.580*	0.706	0.664*	0.684
ACI	0.610	0.609	0.638	0.638	0.758	0.710	0.730

N = 300	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Ex. 5	Ex. 6	Ex. 7
	NR	NNR	NR	NNR	NR	R	R
CPB	0.363	0.363	0.399	0.399	0.482	0.447	0.463
PPE	0.361	0.361	0.400	0.400	0.483	0.443*	0.460
ST	0.359	0.359	0.400	0.400	0.489	0.464*	0.473
ACI	0.402	0.403	0.428	0.428	0.511	0.477	0.497

Table 3.48: Monte Carlo estimates of coverage probabilities of confidence intervals for $-\beta^*_{1,1,0} + \beta^*_{1,1,1}$ (effect of action for history = -1) at the 95% nominal level. Generative models have three stages and two actions per stage. Estimates are constructed using 1000 datasets of size 150, 300 are drawn from each model, and 1000 bootstraps drawn from each dataset. Estimates significantly below 0.95 at the 0.05 level are marked with $*$. Models are designated NR = non-regular, NNR = near-non-regular, R = regular.

CHAPTER IV

Future work

This thesis proposed a framework for constructing adaptive confidence intervals for non-regular functionals which arise in the context of machine learning problems. In this chapter, we briefly discuss two extensions of this work.

4.1 Inference after model selection via the Lasso

Since the seminal paper of Tibshirani (1996) on the Lasso, there has been an explosion of interest in regularization methods that lead to automatic variable selection. The literature is so vast that any attempt to cover it here fully is hopeless.[1] An abbreviated list of such methods includes Bridge Estimation (Frank and Freedman 1993), Smoothly Clipped Absolute Deviation (Fan and Li 1996), False Discovery Rate Penalization (Abramovich et al. 2000), the Elastic Net (Zou and Hastie 2005, and the Dantzig Selector (Candes and Tao 2005). However, despite fervent interest in these methods, the ability to perform valid statistical inference using these approaches remains elusive. The problem of inference after model selection is a longstanding and well known problem (Brown 1967; Olshen 1973; Freedman 1981, 1983; Breiman 2001).

[1] At the time of this writing, Google Scholar reports more than three-thousand citations of Tibshirani's 1996 paper.

The serious pitfalls of performing unadjusted inference after model selection is eloquently and poignantly argued in a series of papers by Leeb and his coauthors (Leeb, Benedikt, and Potscher 1999, 2000, 2003, 2005; Kabaila and Leeb 2006).

A direction for future application of the adaptive confidence interval framework developed in this thesis is inference after model selection. To illustrate this idea, we consider inference for predictions using a linear model fit via the Lasso. Consider the following generative model

$$Y_i = x_i^\intercal \beta^* + \epsilon_i,$$

where we assume that $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are *iid* with mean zero and variance σ^2. We assume that the covariates have been scaled and centered. Furthermore, without loss of generality, we assume that $\bar{Y} \triangleq \frac{1}{n}\sum_{i=1}^n Y_i = 0$. Suppose unknown parameter β^* is estimated usingthe value $\hat{\beta}_n$ which minimizes an L_1 penalized least squares criterion, that is

$$\hat{\beta}_n \triangleq \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (Y_i - x_i^\intercal \beta)^2 + \alpha_n \sum_{j=1}^p |\beta_j|.$$

Let c be a fixed vector in $\mathbb{R}^p$, we consider the problem of constructing an asymptotically valid confidence interval for $c^\intercal \beta^*$. The vector c may represent some contrast of scientific interest, the gradient of a smooth non-linear function, or a future unlabeled input.

It is well known (see Knight and Fu 2000), that the usual starting point for inference, $\sqrt{n}(\hat{\beta}_n - \beta^*)$ is non-regular if $\beta_j^* = 0$ for any $j = 1, 2, \ldots, p$. Moreover, the degree of non-regularity is proportional to number of zero components of β^* (more on this below). Thus, standard statistical approaches to statistical inference like the bootstrap or a Taylor series approximation may not perform well under certain generative models. In this sub-chapter, we propose smooth, adaptive, and data-dependent upper and lower bounds on $c^\intercal \sqrt{n}(\hat{\beta}_n - \beta^*)$ that can be consistently bootstrapped to form confidence intervals. This problem is qualitatively difference from the classification and Q-learning examples presented in previous chapters of this thesis because $c^\intercal\sqrt{n}(\hat{\beta}_n - \beta^*)$ is only defined *implicitly* as the minimizer of a random process. Thus, the above bounds must be based upon this random process rather than an explicit form for $c^\intercal\sqrt{n}(\hat{\beta}_n - \beta^*)$. It will be useful at this point to formalize the problem and consider a simple example.

4.1.1 Non-regularity of the Lasso

Following Knight and Fu (2000) we will make use of the following assumptions.

(A1) Define $\Omega_n \triangleq \frac{1}{n}\sum_{i=1}^n x_i x_i^\intercal$, then $\Omega_n \to \Omega$, and Ω is positive definite.

(A2) $\frac{1}{n}\max_{i\le n} ||x_i||^2 \to 0$.

(A3) $\alpha_n/\sqrt{n} \to \alpha_0 \ge 0$.

(B1) $Y_{ni} = x_{ni}^\intercal(\beta^* + t/\sqrt{n}) + \epsilon_{ni}$, where t is fixed, and ϵ_{ni} are *iid* with mean zero and variance σ^2.

(B2) Define $\Psi_n \triangleq \frac{1}{n}\sum_{i=1}^n x_{ni} x_{ni}^\intercal$, then $\Psi_n \to \Omega$, and Ω is positive definite.

(B3) $\frac{1}{n}\max_{i\le n} ||x_{ni}||^2 \to 0$.

These conditions are standard in least squares problems. In the least squares context, the first condition ensures uniqueness of β^*, and the second is a sufficient condition for asymptotic normality (via the Hajek-Sidak CLT, see DasGupta 2008). Assumption (A3) regards a user-chosen tuning parameter and thus can always be satisfied. Assumptions (B1)-(B3) are to facilitate a discussion of regularity and local properties of both the Lasso and the bounds developed below.

The quantity $\sqrt{n}(\hat{\beta}_n - \beta^*)$ is defined implicitly as the minimizer of the random process $V_n(u)$ where

$$V_n(u) \triangleq \sum_{i=1}^{n} \left(\left(Y_i - x_i^\intercal(\beta^* + u/\sqrt{n})\right)^2 - \epsilon_i^2 \right) + \alpha_n \sum_{j=1}^{p} \left(|\beta_j^* + u_j/\sqrt{n}| - |\beta_j^*| \right).$$

To see this, notice that the above is the localized process where we have re-expressed β as $\beta^* + u/\sqrt{n}$. Thus, the development of an upper bound, say $\mathcal{U}(c)$, on $c^\intercal\sqrt{n}(\hat{\beta}_n - \beta^*)$ should depend on the process $V_n(u)$. The non-smoothness of the L_1 penalty induces non-regularity in $V_n(u)$ and hence in $\sqrt{n}(\hat{\beta}_n - \beta^*)$. Some properties of $\mathbb{V}_n(u)$ are summarized in the following pair of lemmas. Let $\mathbb{Z}$ be p-variate Gaussian distribution with mean zero and covariance Ω.

Lemma 4.1.1 (Knight and Fu 2000)**.** *Assume (A1)-(A3). Then,*

$$V_n(u) \rightsquigarrow V_\infty(u) = -2u^\intercal\mathbb{Z} + u^\intercal\Omega u + \alpha_0 \sum_{j=1}^{p} \left[u_j sign(\beta_j^*) 1_{\beta_j^* \neq 0} + |u_j| 1_{\beta_j = 0} \right]$$

uniformly for u in compact sets.

Lemma 4.1.2 (Knight and Fu 2000)**.** *Assume (A3), and (B1)-(B3). Then,*

$$V_n(u) \rightsquigarrow V_{t,\infty}(u) \triangleq -2u^\intercal\mathbb{Z} + u^\intercal\Omega u + \alpha_0 \sum_{j=1}^{p} \left[u_j sign(\beta_j^*) 1_{\beta_j^* \neq 0} + \left(|u_j + t_j| - |t_j|\right) 1_{\beta_j = 0} \right]$$

uniformly over u in compact sets.

The preceding lemma shows that the limiting distribution of $V_n(u)$ is non-regular. Furthermore, the amount of non-regularity, measured as the number of local parameters that appear in the limiting distribution of $V_n(u)$, is proportional to the number of zero components of β^*. We now illustrate the effect of this non-regularity on the small sample coverage of a residual bootstrap confidence interval using a toy example.

Consider the following generative family of generative models

$$\begin{aligned} \epsilon_i &\sim_{iid} N(0, 1/4), \\ X_i &\sim_{iid} N_p(0, I_p), \\ \beta_j^* &\triangleq 1_{j \leq 5}, \end{aligned} \tag{4.1}$$

where $p \geq 5$ denotes the dimension of the model. Notice that the amount of non-regularity is given by $p - 5$ and is thus increasing in p. We will keep the training set size fixed at $n = 100$ and vary p. For each of 1000 Monte Carlo iterations, and for each $p = 5, 10, \ldots, 50$, we perform the following procedure

1. Draw a value c from $N_p(0, I_p)$.
2. Draw a training set of size $n = 100$ using the generative model in (4.1).
3. Fit the Lasso model choosing α_n as the minimizer of cross-validated prediction error. Record the residuals of this fit.
4. Using 1000 bootstrap resamples of the residuals formed in the preceding step, form a 95% two-sided residual bootstrap confidence interval for $c^\intercal \beta^*$ (see Efron and Tibshirani 2000).
5. Record whether or not the constructed confidence interval covered $c^\intercal \beta^*$.

Averaging over all 1000 Monte Carlo iterations gives an estimate of the achieved coverage probability for the residual bootstrap at each value of p. Figure (4.1.1) shows the estimated coverage for the residual bootstrap. As might be expected from the relationship between the degree of non-regularity and the dimension p of the generative model, the performance of the residual bootstrap deteriorates severely as p increases. When $p = 50$ the estimated coverage is nearly 15% below nominal coverage. The figure also shows the coverage of the adaptive confidence interval which we discuss in the next section.

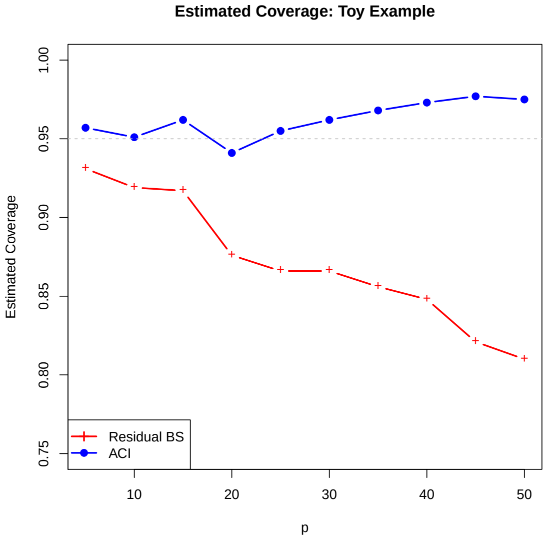

Figure 4.1: Estimated coverage of the residual bootstrap (Residual BS) and the adaptive confidence interval (ACI). Target coverage is 95%. Estimates are based on 1000 Monte Carlo iterations and 1000 bootstrap resamples.

4.1.2 Adaptive confidence intervals for the Lasso

In this section we propose a data-dependent upper bound for $c^{\intercal}\sqrt{n}(\hat{\beta}_n - \beta^*)$. Let σ_j^2 denote the asymptotic variance of $\hat{\beta}_j$. Define the random process

$$\begin{aligned}\Upsilon_{t,n}(u) \triangleq \sum_{i=1}^{n}\left(\left(Y_i - x_i^{\intercal}\left(\beta^* + u/\sqrt{n}\right)\right)^2 - \epsilon_i^2\right) \\ + \frac{\alpha_n}{\sqrt{n}}\sum_{j=1}^{p}\Bigg(\left(|\sqrt{n}\beta_j^* + t_j + u_j| - |\sqrt{n}\beta_j^* + t_j|\right) 1_{\frac{n(\hat{\beta}_{n,j})^2}{\sigma_j^2}\le\lambda_n} \\ + \left(|\sqrt{n}\beta_j^* + u_j| - |\sqrt{n}\beta_j^*|\right) 1_{\frac{n(\hat{\beta}_{n,j})^2}{\sigma_j^2}>\lambda_n}\Bigg).\end{aligned}$$

Notice that $\Upsilon_{0,n}(u) = V_n(u)$, and for any fixed t the $\Upsilon_{t,n}(u)$ is convex and has a unique minimum (Osborne et al. 2000). Intuitively, $\Upsilon_{t,n}(u)$ is a generalized version of $\mathbb{V}_n(u)$ where the term corresponding to the L_1 penalty has been divided into two pieces (i) one applying to indices j where a preliminary test has determined that β_j^* is non-zero, and (ii) one applying to indices j where a preliminary test as failed to determine β_j^* is non-zero. The penalty corresponding indices associated with a failure to reject the hypothesis $\beta_j^* = 0$, have an additional term, t_j, this extra term can be viewed as a local perturbation of $\sqrt{n}\beta_j^*$. The upper bound is constructed by taking a supremum over all minimizers of the locally perturbed process. For any fixed vector $c \in \mathbb{R}^p$ define the upper bound $\mathcal{U}(c)$ as

$$\mathcal{U}(c) \triangleq \sup_{t \in \mathcal{S}_n} c^{\intercal} \left(\arg\min_{u \in \mathbb{R}^p} \Upsilon_{t,n}(u) \right), \tag{4.2}$$

where $\mathcal{S}_n$ is some, possibly data-dependent, subset of $\mathbb{R}^p$. If zero belongs to $\mathbb{R}^p$, then it follows that $c^{\intercal}\sqrt{n}(\hat{\beta}_n - \beta^*) \leq \mathcal{U}(c)$. A lower bound, say $\mathcal{L}(c)$, on $c^{\intercal}\sqrt{n}(\hat{\beta}_n - \beta^)$ can be constructed by replacing the sup in (4.2) with an inf. The following conjectures constitute future work.

Conjecture 1. *Assume (A1)-(A3) and that λ_n tends to ∞ and satisfies $\lambda_n = o(n)$. Assume $S_n \triangleq \{t \in \mathbb{R}^p : ||t||_\infty^2 \leq \lambda_n\}$, and let c be a fixed vector in $\mathbb{R}^p$. Let $V_{t,\infty}(u)$ be as defined in Lemma (4.1.2). Then,*

$$\mathcal{U}(c) \rightsquigarrow \sup_{t \in \mathbb{R}^p} c^{\intercal} \arg\min_{u \in \mathbb{R}^p} V_{t,\infty}(u).$$

Conjecture 2. *Assume (B1)-(B3) and that λ_n tends to ∞ and satisfies $\lambda_n = o(n)$. Assume $S_n \triangleq \{t \in \mathbb{R}^p : ||t||_\infty^2 \leq \lambda_n\}$, and let c be a fixed vector in $\mathbb{R}^p$. Let $V_{t,\infty}(u)$ be* *as defined in Lemma (4.1.2). Then,*

$$\mathcal{U}(c) \rightsquigarrow \sup_{t\in\mathbb{R}^p} c^\intercal \arg\min_{u\in\mathbb{R}^p} V_{t,\infty}(u).$$

Conjecture 3. *Assume (A1)-(A3) and that λ_n tends to ∞ and satisfies $\lambda_n = o(n)$. Assume $S_n \triangleq \{t \in \mathbb{R}^p : ||t||_\infty^2 \leq \lambda_n\}$, and let c be a fixed vector in $\mathbb{R}^p$. Furthermore, let $\mathcal{U}^{(b)}(c)$ denote the bootstrap analog of $\mathcal{U}(c)$. Then, $\mathcal{U}(c)$ and $\mathcal{U}^{(b)}(c)$ have the same limiting distribution in probability. That is,*

$$\sup_{\nu\in BL_1((R)}\left|\mathbb{E}_{\mathcal{M}}\nu\left(\mathcal{U}^{(b)}(c)\right) - \mathbb{E}\nu\left(\mathcal{U}(c)\right)\right|$$

converges to zero in probability.

The first conjecture above concerns the adaptivity of the ACI to the underlying non-regularity. If $\beta_j^* \neq 0$ for all $j = 1, 2, \ldots, p$ then $V_{t,\infty}(u)$ does not depend on t and equals $V_\infty(u)$. Thus, in this case, $\mathcal{U}(c)$ and $c^\intercal\sqrt{n}(\hat{\beta}_n - \beta^*)$ have the same limiting distribution. The second conjecture states that the upper bound $\mathcal{U}(c)$ is regular. In practice, the bootstrap distribution of the bounds $\mathcal{U}(c)$ and $\mathcal{L}(c)$ are used to form confidence sets. The third conjecture states that the upper and lower bounds can be consistently bootstrapped. Proofs of these results are ongoing work.

The performance of the ACI on the toy example is illustrated in figure (4.1.1). The confidence interval was formed using $\mathcal{S}_n$ as described in the above conjectures with $\lambda_n = \log\log n$. The ACI delivers nominal coverage for every dimension p but appears to become more conservative as the degree of non-regularity increases. This is somewhat of a concern given the strong desire to apply the Lasso in high dimensional settings. In this same vein, note that the conjectures above are stated for the low (fixed p) setting, future work will include an analysis of the ACI in an asymptotic framework that allows p to diverge with n.

4.1.3 Computation

One of the challenging aspects of the ACI for the Lasso is computation. In order to compute a bootstrap confidence interval one must compute $\mathcal{U}^{(b)}(c)$ for *every* bootstrap sample. For every fixed value of t, computing the arg min of $\Upsilon_{t,n}^{(b)}(u)$ is a quadratic programming problem (see, for example, Osborne, Presnell, and Turlach 2000) and hence can be done efficiently and exactly. A natural first approach to computing $\mathcal{U}^{(b)}(c)$ is to use an iterative search over t. Properties of an iterative optimization approach may be difficult to characterize because properties of $\arg\min_{u\in\mathbb{R}^p} \Upsilon_{t,n}^{(b)}$, viewed as a function of t, are currently not well understood. However, we show below that it is possible to reformulate the computation of $\mathcal{U}^{(b)}(c)$ as a linear mixed integer program which can be solved exactly (that is, to machine precision) using specialized software (e.g. CPLEX). While solving an integer program can be quite computationally burdensome, especially in high-dimensional problems, the integer formulation is useful for a number of reasons. First, the integer program yields an essentially exact solution and thus can be used to benchmark the quality of other, faster, numerical procedures. Second, the form of the integer program can inform the development of relaxed versions of the problem potentially leading to better and more specialized optimization routines. Lastly, many numerical optimization procedures require the choice of one or more tuning parameters which may govern, among other things, the balance of exploration and exploitation and the stopping criteria. Solutions to the integer program can be used to tune these parameters. For example, suppose that an iterative optimization procedure requires the choice of tuning parameter M which denotes the maximum number of steps. Let $\hat{\mathcal{U}}_M^{(b)}(c)$ denote the approximation of $\mathcal{U}^{(b)}(c)$ for the bth bootstrap sample using M as the maximum number of steps.

Let r denote a small number of 'pilot' iterations. Using the integer program, compute $\mathcal{U}^{(1)}(c)$,$\mathcal{U}^{(2)}(c), \ldots$,$\mathcal{U}^{(r)}(c)$. Then, for each $k = 1, 2, \ldots, r$ let M_k denote the smallest number of steps M for which $\hat{\mathcal{U}}_M^{(k)}(c) \geq \mathcal{U}^{(k)}(c) + \epsilon$, where ϵ denotes some acceptable loss in solution quality. Then, one can choose $M = \max_{1 \leq k \leq r} M_k$ as the maximum number of iterations for the remaining $B - r$ iterations needed to compute the bootstrap confidence interval.

We have argued that having a computationally expensive but exact mixed integer program formulation of $\mathcal{U}^{(b)}(c)$ may be beneficial for a number of reasons. We now give a very brief sketch as to how such a formulation is derived. For each fixed t, $\Upsilon_{t,n}^{(b)}(u)$ is convex and in the absence of perfect collineararity, possesses a unique solution (Osborne et al., 2000). Furthermore, since minimizing $\Upsilon_{t,n}^{(b)}(b)$ is equivalent to the minimization problem required by the Lasso, one has strong duality (Osborn et al., 2000) and $\arg\min_{u \in \mathbb{R}^p} \Upsilon_{t,n}(u)$ is completely characterized by the Karush-Kuhn-Tucker (KKT) conditions. Let $\mathrm{KKT}(t, u, \theta)$ denote the KKT conditions for fixed value t, where θ denotes additional slack and dual variables. Computing $\mathcal{U}^{(b)}(c)$ can be re-written as $\sup_{t,u,\theta} c^T u$ subject to the constraints given in $\mathrm{KKT}(t, u, \theta)$ and the requirement that $t \in \mathcal{S}_n$. The KKT conditions are linear constraints with an integer component (see Boyd and Vandenberghe 2004), and provided the set $\mathcal{S}_n$ imposes linear constraints on t, the newly formulated problem is indeed a linear mixed integer program.

4.1.4 Conclusions

The problem of constructing confidence intervals post-model selection is a longstanding and important problem. Preliminary analysis suggests that an adaptive confidence interval may have promise in the case where model selection has been done using the Lasso. However, a number of theoretical and computational issues are directions for future work. This problem is also important in the broader context of developing a unified framework for constructing adaptive confidence intervals in non-regular problems. In particular, the functional of interest in this problem is only defined implicitly as the minimizer of a random processes, forcing the construction of bounds to occur indirectly through the random process. In addition, the Lasso is often employed in the so-called 'large p small n' paradigm, thus, it is imperative to develop theoretical properties of the adaptive confidence interval in a framework that allows for the dimension p to diverge with sample size n.

4.2 Adaptive confidence intervals for the value of a learned DTR

In the second part of this thesis we addressed the problem of constructing confidence intervals for coefficients in a learned dynamic treatment regime (DTR). Another quantity of interest is the expected return, averaged over future patients, of the learned DTR. This quantity is known as the value of a learned DTR and is very closely related to the test error in classification. In this subchapter we briefly discuss the problem of constructing an adaptive confidence interval for the value of a learned DTR in the simplest possible setting, where there is one-stage of treatment and two possible treatments available at each stage.

The setup is as follows. We observe a training set $\mathcal{D} \triangleq \{(Y_i, H_i, A_i)\}_{i=1}^{n}$ drawn *iid* from fixed but unknown distribution P. Each triple (Y, H, A) is composed of a response Y which is coded so that higher values of Y correspond to better clinical outcomes, a vector of covariates H which are assumed to belong to $\mathbb{R}^p$, and a binary treatment A which is coded to take values in $\{-1, 1\}$. Using the training data $\mathcal{D}$, we estimate the optimal DTR using the Q-learning algorithm, as described in Chapter

3. When there is only one stage of treatment,Q-learning reduces to ordinary linear regression. Let $Q(h, a)$ denote the conditional mean of the response Y given treatment history $H = h$ and treatment $A = a$. That is, $Q(h, a) = \mathbb{E}(Y|H = h, A = a)$. We will model $Q(h, a)$ using the following working model

$$Q(h, a; \beta) \triangleq \beta_0^\intercal + a\beta_1^\intercal h_1,$$

where $h_0 = \Psi_0(h)$ and $h_1 = \Psi_1(h)$ are features constructed from $H = h$. Let $\mathbb{P}_n$ denote the empirical measure and let $\hat{\beta}$ denote the least squares estimator

$$\hat{\beta} \triangleq \arg \min_{\beta \in \mathbb{R}^p} \mathbb{P}_n(Y - Q(H, A))^2.$$

Let $\hat{\pi}$ denote the estimated optimal DTR, that is, $\hat{\pi}$ is the map from $\mathbb{R}^p$ into $\{-1, 1\}$ given by

$$\hat{\pi}(h) = \arg \max_{a \in \{-1,1\}} Q(h, a; \hat{\beta}) = sign\left(\hat{\beta}_1^\intercal h_1\right).$$

See Chapter 3 for additional details on Q-learning. The value $V^{\hat{\pi}}$ of the learned policy $\hat{\pi}$ is the expected response Y when treatment for a patient with history $H = h$ is (with probability one) assigned treatment $\hat{\pi}(h)$. That is,

$$V^{\hat{\pi}} \triangleq P^{\hat{\pi}}Y = P\left(\frac{Y}{p(A|H)} 1_{A=\hat{\pi}(H)}\right) = P\left(\frac{Y}{p(A|H)} 1_{AH_1^\intercal \hat{\beta}_1 < 0}\right), \tag{4.3}$$

where $p(a|h) \triangleq P(A = a|H = h)$, and we have assumed that for each a, $p(a|h) > 0$ with (H) probability one. The form of the last term in (4.3) shows that the value can be viewed as a weighted test error. The distribution of the weights influence the degree of regularity. For example if $Y/P(A|H) \approx 0$ with high probability whenever

$H^{\intercal}\beta^{*} = 0$ then

$$P\left(\frac{Y}{p(A|H)}1_{AH_1^{\intercal}\hat{\beta}_1<0}\right) \approx P\left(\frac{Y}{p(A|H)}1_{AH_1^{\intercal}\hat{\beta}_1<0}1_{H^{\intercal}\beta^{*}\neq 0}\right),$$

where the right hand side of the above display is regular. On the other hand, if $\big|Y/P(A|H)\big| \gg 0$ with non-trivial probability whenever $H^{\intercal}\beta^{*} = 0$, the non-regularity will be exacerbated.

Forming a confidence interval for the value function of a learned policy is conceptually and theoretically similar to the problem of constructing a confidence interval for the test error in classification. However, the small sample performance may depend heavily on properties of the weights $Y/p(A|X)$. Furthermore, extending the adaptive confidence interval framework to multistage case will be non-trivial since one must account not only for the non-regularity introduced by the indicator function, but also the non-regularity of the coefficients indexing the regression functions at later stages (see Chapter 3).

BIBLIOGRAPHY

BIBLIOGRAPHY

Donald W. Andrews and Gustavo Soares. Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection. *SSRN eLibrary*, 2007.

Donald W. K. Andrews and Patrik Guggenberger. Validity of subsampling and plug-in asymptotic inference for parameters defined by moment inequalities. *Econometric Theory*, 25(03):669–709, 2009.

D.W.K. Andrews and P. Guggenberger. The limit of finite sample size and a problem with subsampling and with the m out of n bootstrap. *Cowles Foundation*, 2005a.

D.W.K. Andrews and P. Guggenberger. Hybrid and size-corrected subsample methods. *Cowles Foundation Discussion Paper No. 1606*, 2005b.

D.W.K. Andrews and P. Guggenberger. Invalidity of subsampling in- ference based on post-consistent model selection estimators. *Cowles Foundation*, 2005c.

D.W.K. Andrews and P. Guggenberger. Applications of subsampling, hybrid, and size-correction methods. *Cowles Foundation Discussion Paper No. 1608*, 2005d.

Martin M. Anthony and Peter Bartlett. *Learning in Neural Networks: Theoretical Foundations*. Cambridge University Press, New York, NY, USA, 1999. ISBN 052157353X.

K. B. Athreya. Bootstrap of the mean in the infinite variance case. *The Annals of Statistics*, 15(2):pp. 724–731, 1987. ISSN 00905364. URL `http://www.jstor.org/stable/2241336`.

P. Bartlett, M. Jordan, and J. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistician*, 101:138–156, 2005. URL `citeseer.ist.psu.edu/628679.html`.

P. J. Bickel, F. Gotze, and W. R. Van Zwet. Resampling fewer than n observations: gains, losses, and remedies for losses. statistica sinica 7. *Statist. Sinica*, 7:1–32, 1997.

PJ Bickel, AJ Klaassen, Y. Ritov, and JA Wellner. Efficient and adaptive inference in semi-parametric models, 1993.

D. Blatt, Susan A. Murphy, and J. Zhu. A-learning for approximation planning, 2004.

Arup Bose and Snigdhansu Chatterjee. Generalized bootstrap for estimators of minimizers of convex functionals. Technical report, Indian Statistical Institute, 2000.

Arup Bose and Snigdhansu Chatterjee. Generalized bootstrap for estimators of minimizers of convex functions. *Journal of Statistical Planning and Inference*, 117(2): 225 – 239, 2003. ISSN 0378-3758. doi: DOI: 10.1016/S0378-3758(02)00386-5. URL `http://www.sciencedirect.com/science/article/B6V0M-472JW2H-3H/2/6e1a2a0985d57c01`

J.Z. Bretagnolle. Lois limites du bootstrap de certaines fonctionnelles. *Annals de L'institut Henri Poincar*, 19:281–296, 1983.

B. Chakraborty, S. Murphy, and V. Strecher. Inference for non-regular parameters in optimal dynamic treatment regimes. *Statistical Methods in Medical Research*, page 0962280209105013v1, 2009.

A Chatterjee and S.N. Lahiri. Bootstrapping Lasso estimators. *Peronsal communication.*, 2009.

Xu Cheng. Robust confidence intervals in nonlinear regression under weak identification. *Job Market Paper*, 2008.

M.C. Chernick, V.K. Murthy, and C.D. Nealy. Application of bootstrap and other resampling techniques: Evaluation of classifier performance. *PRL*, 3:167–178, 1985.

Hie-Choon Chung and Chien-Pai Han. Conditional confidence intervals for classification error rate. *Computational Statistics and Data Analysis*, 53:4358–4369, 2009.

Sándor Csörgő and Andrew Rosalsky. A survey of limit laws for bootstrapped sums. *International Journal of Mathematics and Mathematical Statistics*, 45:2835–2861, 2003.

Donald W.K. Donald. Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, 69:683–734, 2001.

Hani Doss and Jayaram Sethuraman. The price of bias reduction when there is no unbiased estimate. *The Annals of Statistics*, 17(1):pp. 440–442, 1989. ISSN 00905364. URL `http://www.jstor.org/stable/2241528`.

Lutz Dmbgen. On nondifferentiable functions and the bootstrap. *Probability Theory and Related Fields*, 95:125–140, 1993. ISSN 0178-8051. URL `http://dx.doi.org/10.1007/BF01197342`. 10.1007/BF01197342.

RB DAgostino. Departures from normality, tests for. *Encyclopedia of Statistical Sciences*, 2:315–324, 1982.

Bradley Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983. doi: 10.2307/2288636. URL `http://dx.doi.org/10.2307/2288636`.

Bradley Efron and Robert Tibshirani. Cross-validation and the bootstrap: Estimating the error rate of a prediction rule. Technical Report 172, Stanford, 1995.

Bradley Efron and Robert Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560, 1997. URL `http://www.jstor.org/stable/2965703`.

S.J. Haberman. Concavity and estimation. *Annals of Statistics*, 17:1631–1661, 1989.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2009.

R. Henderson, P. Ansell, and D. Alshibani. Regret-Regression for Optimal Dynamic Treatment Regimes. *Biometrics*, 9999(9999), 2009.

Keisuke Hirano and Jack Porter. Impossibility results for nondifferentiable functionals. 2009. URL `http://econpapers.repec.org/RePEc:pra:mprapa:15990`.

A. Isaksson, M. Wallman, H. Gransson, and M.G. Gustafsson. Cross-validation and bootstrapping are unreliable in small sample classification. *Pattern Recognition Letters*, 29(14):1960 – 1965, 2008. ISSN 0167-8655. doi: DOI: 10.1016/j.patrec.2008.06.018. URL `http://www.sciencedirect.com/science/article/B6V15-4T0FF18-1/2/cccdbb72b6a9a57c8`

W. Jiang, S. Varma, and R. Simon. Calculating confidence intervals for prediction error in microarray classification using resampling. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.

Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, pages 1137–1145. Morgan Kaufmann, 1995.

Michael R. Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer, 2008a.

M.R. Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Verlag, 2008b.

W.J. Krzanowski and D.J. Hand. Assessing error rate estimators: The leave-one-out method reconsidered. *PRL*, 3:167–178, 1985.

Eric B Laber and Susan A Murphy. Small sample inference for generalization error in classification using the cud bound. In *Proceedings of the Proceedings of the Twenty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, pages 357–365, Corvallis, Oregon, 2008. AUAI Press.

Eric B Laber and Susan A Murphy. Adaptive confidence intervals for the test error in classification. Technical Report 497, University of Michigan, 2009.

J.A. Lieberman, T.S. Stroup, J.P. McEvoy, M.S. Swartz, R.A. Rosenheck, D.O. Perkins, R.S.E. Keefe, S.M. Davis, C.E. Davis, B.D. Lebowitz, et al. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *The New England journal of medicine*, 353(12):1209, 2005.

E.E.M. Moodie, T.S. Richardson, and D.A. Stephens. Estimating optimal dynamic regimes: Correcting bias under the null. *Biometrics*, 63(2):447–455, 2010.

Erica E. Moodie, Thomas S. Richardson, and David A. Stephens. Demystifying optimal dynamic treatment regimes, 2005.

Susan A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B*, 65(2):331–366, 2002.

Susan A. Murphy. A generalization error for q-learning. *Journal of Machine Learning Research*, 6:1073–1097, Jul 2005.

W. Niemiro. Asymptotics for m-estimators defined by convex minimization. *Annals of Statistics*, 20:1514–1533, 1992.

R.A. Olshen. The conditional level of the F-test. *Journal of the American Statistical Association*, 68(343):692–698, 1973.

W.E. Pelham Jr and G.A. Fabiano. Evidence-based psychosocial treatments for attention-deficit/hyperactivity disorder. *Journal of Clinical Child & Adolescent Psychology*, 37(1):184–214, 2008.

S. Pliszka. AACAP Work Group on Quality Issues. Practice parameter for the assessment and treatment of children and adolescents with attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry*, 46(7):894–921, 2007.

Dimitris N. Politis and Joseph P. Romano. Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, 22(4):pp. 2031–2050, 1994. ISSN 00905364. URL `http://www.jstor.org/stable/2242497`.

Min Qian and Susan A. Murphy. Performance Guarantees for Individualized Treatment Rules. *Technical report number 498*, 2009.

J.M. Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*, 2004.

A.J. Rush, M. Trivedi, and M. Fava. Depression, IV: STAR* D treatment trial for depression. *American Journal of Psychiatry*, 160(2):237, 2003.

Richard Samworth. A note on methods of restoring consistency to the bootstrap. *Biometrika*, 90(4):985–990, 2003. URL `http://econpapers.repec.org/RePEc:oup:biomet:v:90:y:2003:i:4:p:985-990`.

Rosa A. Schiavo and D.J. Hand. Ten more years of error rate research. *International Statistical Review*, 68:295–310, 2000.

Jun Shao. Bootstrap sample size in nonregular cases. *Proceedings of the American Mathematical Society*, 122(4):pp. 1251–1262, 1994. ISSN 00029939. URL `http://www.jstor.org/stable/2161196`.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):pp. 267–288, 1996. ISSN 00359246. URL `http://www.jstor.org/stable/2346178`.

Aad van der Vaart. On differentiable functionals. *Annals of Statistics*, 19(1):178–204, 1991.

AW Van der Vaart and J.A. Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Verlag, 1996.

C.J.C.H. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3):279–292, 1992.

Minge Xie, Kesar Singh, and Cun-Hui Zhang. Confidence intervals for population ranks in the presence of ties and near ties. *Journal of the American Statistical Association*, 104(486):775–788, 2009. doi: 10.1198/jasa.2009.0142. URL `http://pubs.amstat.org/doi/abs/10.1198/jasa.2009.0142`.

Yuhong Yang. Comparing learning methods for classification. *Statistica Sinica*, 16: 635–657, 2006.

P. Zhang. Ape and models for categorical panel data. *Scandinavian Journal of Statistics*, (22):83–94, 1995.

Y. Zhao, M.R. Kosorok, and D. Zeng. Reinforcement learning design for cancer clinical trials. *Statistics in Medicine*, 28:3294–3315, 2009.