UNIVERSITY OF CALIFORNIA SAN DIEGO

Essays in Econometrics

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Economics

by

Daniel Pellatt

Committee in charge:

Professor Yixiao Sun, Chair
Professor Richard T. Carson Jr
Professor Dimitris Politis
Professor Allan Timmermann
Professor Kaspar Wuthrich

2023

The Dissertation of Daniel Pellatt is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

To Lila, Mom, Dad, and Andrew.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First I would like to acknowledge my adviser and committee chair Yixiao Sun for his support during my graduate studies at UCSD. I am deeply grateful for his mentorship, the many insights and suggestions he provided throughout my work, and his investment in me as a student and researcher.

I am indebted to other faculty members as well. Kaspar Wuthrich always prioritized the success of myself and other graduate students. I greatly appreciate his helpful advice and thoughtful comments. Allan Timmermann broadened my exposure to different areas of econometrics and his feedback was always insightful. Richard Carson provided valuable viewpoints and pushed my reasoning in productive directions. I am also grateful to Dimitris Politis for providing an additional and valuable cross-disciplinary perspective on my committee. I have been fortunate to interact with an extremely talented group of academics at UCSD who have been gracious with their time and devoted to the success of their students.

On a personal level I am very grateful to Lila for her support and companionship, to my parents who are the foundation of my familial support and are always concerned for my well-being, and to my brother Andrew for his friendship and wine recommendations. Friends including Zach, Roman, Cameron, Nikolay, and Yu-Chang, among others, have also provided support along the way.

Lastly, I am very appreciative of Lajos Horváth at the University of Utah. Lajos takes great care to foster potential in his students and his guidance during my master's studies improved my analytical focus and prepared me for challenges at UCSD and beyond.

Chapter 1 contains material being prepared for submission for academic publication. The dissertation author is the sole author of this material.

Chapter 2 contains material being prepared for submission for academic publication. It is joint work with Yixiao Sun. The dissertation author is a primary author of this material.

Chapter 3, in full, is a reprint of the material as it appears in Asymptotic F Test in

Regressions With Observations Collected at High Frequency Over Long Span 2022. Pellatt, Daniel F.; Sun, Yixiao, Journal of Econometrics, 2022. Minor adjustments around the referencing and title of an appendix have been made to integrate the format with that of this dissertation. The dissertation author is a primary author of this material.

VITA

2011	Bachelors of Science, University of Oregon
2015	Masters of Statistics, University of Utah
2023	Doctor of Philosophy, University of California San Diego

ABSTRACT OF THE DISSERTATION

Essays in Econometrics

by

Daniel Pellatt

Doctor of Philosophy in Economics

University of California San Diego, 2023

Professor Yixiao Sun, Chair

Each chapter of this dissertation examines a different econometric problem of interest and proposes a new approach to the data analysis problem at hand. The chapter titles may give the impression that some of these topics lie in disparate areas of focus. The topic and approach of Chapter 3, for example, shares less commonalities with Chapters 1 and 2 than the first two chapters share with one another. A connecting theme between all three chapters is the combination of foundational problems with modern data or methodologies designed to accommodate modern data analysis techniques.

In the first two chapters, the PAC-Bayesian analytical framework, which has developed alongside the growth of machine learning applications, drives analyses of more traditional

problems involving binary decision and individual treatment rules. In Chapter 1, this facilitates the derivation of new individual treatment rule estimators in the setting where a policy maker faces a general budget or resource constraint. In Chapter 2, this suggests new decision rules when a policy maker has a general utility function over payoffs that may have asymmetries and vary with covariates relevant to the decision problem. In each case the rules possess desirable theoretical properties, perform competitively against state-of-the-art alternatives, and have additional advantages in terms of applicability, estimation options, and modeling flexibility.

Chapter 3 considers hypothesis testing in linear regressions when observations may be sampled at short time intervals. Whereas monthly or even quarterly observations were once ubiquitous in time series regression applications, it is becoming more common to have weekly, daily or even intraday observations. However, higher frequency data can pose challenges for classical inference procedures. F tests are proposed that utilize series long run variance estimation. Under reasonable discrete-time or continuous-time settings, the procedures yield valid inference so that the proposed hypothesis tests are robust to the sampling interval available to the practitioner. The tests have competitive size and power properties against the limited set of alternatives in a simulation study. Finally, an empirical example examining a relationship between interest rates associated with shorter and longer duration bonds illustrates the usefulness of the procedure.

Chapter 1

PAC-Bayesian Treatment Allocation Under Budget Constraints

Abstract

This paper considers the estimation of treatment assignment rules when the policy maker faces a general budget or resource constraint. Utilizing the PAC-Bayesian framework, we propose new treatment assignment rules that allow for flexible notions of treatment outcome, treatment cost, and a budget constraint. For example, the constraint setting allows for cost-savings, when the costs of non-treatment exceed those of treatment for a subpopulation, to be factored into the budget. It also accommodates simpler settings, such as quantity constraints, and doesn't require outcome responses and costs to have the same unit of measurement. Importantly, the approach accounts for settings where budget or resource limitations may preclude treating all that can benefit, where costs may vary with individual characteristics, and where there may be uncertainty regarding the cost of treatment rules of interest. Despite the nomenclature, our theoretical analysis examines frequentist properties of the proposed rules. For stochastic rules that typically approach budget-penalized empirical welfare maximizing policies in larger samples, we derive non-asymptotic generalization bounds for the target population costs and sharp oracle-type inequalities that compare the rules' welfare regret to that of optimal policies in relevant budget categories. A closely related, non-stochastic, model aggregation treatment assignment rule is shown to inherit desirable attributes.

1.1 Introduction

This paper proposes new statistical decision rules for treatment assignments under a general budget or resource constraint. A key objective in the empirical analysis of treatment data is identifying policies that result in the most beneficial outcomes. There is a large literature (e.g. Manski (2004) and Hirano and Porter (2009)) that examines how to determine which policies are optimal to implement in the absence of constraints such as one on policy cost. In practice, however, policy makers are rarely free from constraints when it comes to the policies they may enact. Several recent papers in the econometrics literature, including Kitagawa and Tetenov (2018), Athey and Wager (2021), and Mbakop and Tabord-Meehan (2021), consider the treatment estimation problem from an empirical welfare maximization (EWM) perspective that allows for arbitrary constraints on the functional form of the decision rule. However, these papers do not address general budget constraints nor cost uncertainty that varies with the characteristics of individual agents. For example, while Kitagawa and Tetenov (2018) consider quantity constraints via random rationing, this treats costs as fixed and hence cannot identify which policies most efficiently balance cost vs. outcome trade-offs when costs vary with individual characteristics.

Here we focus on the setting where costs may be uncertain, current resource limitations may preclude treating all that can benefit, and where individual characteristics can influence treatment responses and costs. Compared to the unconstrained setting, the theoretically optimal treatment rule involves population objects that are more difficult to estimate and analyze in concert. For example, Bhattacharya and Dupas (2012) show that under a quantity constraint, which is simpler than the setting with variable costs, the optimal rule is to assign treatment when the conditional average treatment effect exceeds its $(1-c)$th quantile. Here c is the maximal proportion of treatments assignable under the constraint. As a result, it can be difficult to evaluate properties of interest for proposed approaches and each existing approach has limitations.

The contributions of the paper are as follows. First, we propose new treatment rules that expand the tool set available to policy makers in the budget constrained setting. Second, we show they possess several potential benefits in terms of theoretical guarantees, the variety of settings in which they can be applied, and ease of estimation. Third, we show expert knowledge can be incorporated when the policy maker has non-data-dependent insights into the problem. However, the ability to integrate expert knowledge is a secondary feature of the approach. In our primary implementation we assume no such knowledge.

PAC-Bayesian analysis applies the probably approximately correct learning framework to objects of interest that involve probability distributions over model or parameter families. These objects can include, for example, treatment rules formed by aggregating over a family of potential rules. Our work can be seen as extending the PAC-Bayesian learning approach to the treatment setting in a way that incorporates a secondary cost objective. This motivates the proposed rules and allows us to derive generalization bounds for the costs and oracle-type inequalities for the welfare regret of proposed rules. Here, the welfare regret associated with a treatment rule is the loss in expected welfare of the decision rule relative to the theoretically optimal decision rule (cf. Manski (2004)). To work within the regret framework, we also derive the form of a theoretically optimal treatment policy if the data generating process (DGP) were known under a general budget constraint.

Individualized treatment policies under budget restrictions are of interest in a variety of settings. Often policy makers with limited resources face uncertainty regarding the costs and benefits of potential policies where this uncertainty is driven due to the fact that costs and benefits vary with the individual characteristics of those who decide to participate in a program. For example, Finkelstein et al. (2012) examine outcomes such as health care utilization and self-reported health measures following a randomized expansion of household access to Medicaid in Oregon. A policy maker may be interested in identifying policies to maximize a well-defined weighted average of such outcomes given a binding expenditure constraint. The government has control over eligibility rules defined on characteristics such as age, income, and the number of children in a household that directly influence expected cost and cost uncertainty.

Insecticide-treated nets (ITNs) for protection against malaria in regions of Africa represent another common example. Lengeler (1998), for instance, documents reductions in child mortality while Kuecken et al. (2014) document returns to education related to ITN provisions. Teklehaimanot et al. (2007) estimate the cost of providing an ITN to every at-risk individual in sub-Saharan Africa to be 2.5 billion dollars. However, government and aid funding was below that level at the time of the study. Bhattacharya and Dupas (2012) look at a treatment policy estimator under quantity constraints derived from data from a randomized experiment assigning ITNs to rural households in Kenya. They use fixed costs to estimate rules that satisfy quantity constraints. Our approach makes it possible to target policies in such a way as to account for cost heterogeneity (e.g. different distribution channels) and hence improve efficiency and achieve a higher overall outcome level.

Beyond aid and social safety net policies, the budget constrained treatment assignment problem can also arise in a commercial context for firms considering potentially costly promotions aimed at obtaining new customers. For instance, Sun et al. (2021) recently proposed a budget constrained treatment estimator aimed at determining which customers should be offered trial access to a premium service. They seek to use customers' individual characteristics to discriminate against making offers to customers likely to heavily utilize the service in the trial (high cost) while being unlikely to use the service after the trial period expires. Rather than the simple notion of not wanting to implement a policy that leads to long-term losses, many companies will also face a short-term constraint on how much they can "lose" in the trial phase to gain market share. For other firms, like Uber which is considered in Sun et al. (2021), a deeper issue may arise. Increasing sales or trial offers may fundamentally alter the firm's cost structure (e.g., increasing driver compensation to induce enough new drivers to work to handle the increased number of trips).

The rules we develop start from a user-specified family of (non-stochastic) treatment models $\mathscr{F}$ that map an individual's covariates that are observable pre-treatment to the $\{0,1\}$ treatment indicator space. Rather than choosing the model that maximizes the empirical welfare in $\mathscr{F}$, for example, we instead consider stochastic treatment rules derived from $\mathscr{F}$ and a measure of budget penalized empirical welfare. Given an individual's pre-treatment covariates, their treatment probability is calculated as an exponentially weighted average over the treatments specified by members of $\mathscr{F}$. The treatment probability is similar to a weighted majority vote taken over $\mathscr{F}$. The exponential weighting received by members of the model family is greatest for models with a large budget-penalized empirical welfare. The magnitude of the penalization term related to cost is determined by a parameter u that modulates the trade-off between maximizing welfare and reducing costs. Any choice for u will correspond to a different maximal empirical budget, with $u = 0$ corresponding to an unlimited budget (no constraint). Typically, for larger sample sizes, the rule is unlikely to assign identical covariates to different treatments unless there are subsets of the model family with similarly high values of penalized welfare that prescribe different treatments. We also consider closely related, non-stochastic, model aggregation treatment rules that aggregate over $\mathscr{F}$ to make treatment decisions.

Utilizing a PAC-Bayesian framework, under reasonable conditions we show that for a set of u values, in large samples, with high probability we obtain increasingly accurate estimates of the target population costs associated with corresponding stochastic treatment rules. We can use these estimates to select u or, alternatively, u can be chosen via cross-validation. At the same time, with u chosen in either manner, with high probability the resulting rule achieves a welfare regret comparable to that of the best models in the model family that have a similar target population cost. Starting from a set of budget penalty parameters, the policy maker can trace out good estimates of the feasible target population budgets, select the parameter associated with one of these estimates, and obtain a treatment rule with desirable regret properties. Regarding the non-stochastic, model aggregation treatment rules, we show that they inherit desirable properties from the stochastic rules. We also consider the setting where u is chosen to meet a predetermined target population budget level. The procedure in this case is still reasonably motivated, as the rule minimizes an upper bound on the target population regret among rules that satisfy an empirical budget constraint. However, the generalization bounds for the target population cost and the oracle-type inequalities in this case become more complex to interpret.

The remainder of the paper is organized as follows. Section 1.2 discusses related literature and papers with alternative budget constrained treatment estimators. Section 1.3 details the statistical setting, treatment model formulation, and initial properties useful for later results. Section 1.4 provides theoretical motivation for the proposed treatment rules, utilizing the PAC-Bayesian analysis framework to examine (frequentist) properties of the proposed rules. Section 1.5 conducts a simulation experiment and discusses implementation and estimation. Lastly, Section 1.6 conducts a short empirical illustration utilizing data from the Job Training Partnership Act Study and Section 1.7 concludes.

1.2 Related Literature

The topic of budget constrained treatment allocation is the subject of a small but growing literature. Sun et al. (2021) and Wang et al. (2018) empirically implement treatment rules starting from the notion of a theoretically optimal rule. They estimate unknown population level objects that appear in the optimal rules and then plug in the empirical counterparts to the corresponding theoretical formulas to obtain rules. The standard drawback of this sort of approach is that the estimation technique doesn't directly target policies that maximize the welfare problem of interest. For example, the regressions utilized to fit the conditional average treatment and cost functions in Wang et al. (2018) might yield parameters that are most accurate in regions of the covariate space that are less important for distinguishing individuals with a high outcome-to-cost ratios in the population. Wang et al. (2018) also consider a second method that shares similarities with the approach taken by Huang and Xu (2020). These approaches add the budget constraint to the outcome-weighted treatment learning approach considered, for example, in Zhao et al. (2012). These approaches work from optimization problems that directly target an empirical version of the problem of interest.

One drawback of the aforementioned techniques is a lack of theoretical insight regarding the true target population cost and risk attributes of the proposed rules. Sun (2021) adapts the

EWM setting of Kitagawa and Tetenov (2018) to account for a general budget constraint. She considers a conservative rule that will satisfy the budget constraint asymptotically. She also considers a modified rule where a Lagrange multiplier parameter is capped during estimation. This will, asymptotically, approach the welfare of the budget constrained welfare maximizing policy among the user-specified model class. This methodology extends the arbitrary form features of EWM to the budget constraint setting. However, the rules involve a non-convex estimation procedure that may become difficult if the model class includes more flexible functional forms. While our methodology sacrifices some ability to satisfy functional form constraints due to its stochastic nature, one benefit is that we can take advantage of Bayesian estimation machinery as discussed in Section 1.5. Lastly, although the modified rules of Sun (2021) will approach the optimal rule within the original budget constraint, it is worth noting that the modified rule may violate that budget constraint. One benefit of our approach is we can compare our rules to those with the highest welfare among rules with the same target population cost as the proposed rules.

In a broader context, this paper contributes to a growing literature on statistical treatment rules in econometrics, including Manski (2004), Dehejia (2005), Hirano and Porter (2009), Bhattacharya and Dupas (2012), Kitagawa and Tetenov (2018), Viviano (2019), and Athey and Wager (2021). This literature has overlap with additional fields including statistics and machine learning. For examples, see Qian and Murphy (2011) and Beygelzimer and Langford (2009), respectively. Additional references and a discussion of the links between these fields can be found in Athey and Wager (2021). In the machine learning literature, London and Sandler (2019) utilize a PAC-Bayesian approach to policy estimation for the logged bandit feedback problem which is closely related to treatment policy estimation. We also note that Kitagawa et al. (2023) examine stochastic treatment assignment rules from a PAC-Bayesian perspective. Their paper's approach has overlap with ours, however the papers diverge in a number of dimensions stemming from our focus on the setting with a general budget constraint which is not considered there.

Lastly, our analysis and proposed treatment rules are heavily influenced by the PAC-Bayesian machine learning literature. Seminal works in this area include Shawe-Taylor and

Williamson (1997), McAllester (1999b), McAllester (1999a), Seeger (2002), and McAllester (2003b). In particular, we utilize techniques stemming from Catoni (2007), Lever et al. (2010), Maurer (2004), Germain et al. (2015), and Alquier et al. (2016). The theoretical contribution of our paper is, first, to modify and adapt relevant tools and generalization bounds to the treatment choice setting. We also develop the incorporation of a secondary objective or loss function (the treatment cost cost) into the analysis that yields informative oracle-type inequalities and generalization bounds relevant to the constrained budget setting.

1.3 Setup and Assumptions

1.3.1 Statistical Setting and Policy Maker's Problem

We consider the setting where a policy maker has data consisting of observations

$$Z_i = (Y_i, C_i, D_i, X_i),\ i = 1, \ldots, n.$$

Here, $X_i \in \mathscr{X} \subset \mathbb{R}^{d_x}$, where $d_x \in \mathbb{N}$, denotes a vector of covariates for individual or unit i observed prior to treatment assignment, $Y_i \in \mathbb{R}$ is unit i's outcome that is observed after treatment assignment, $C_i \in \mathbb{R}$ is the cost incurred and $D_i \in \{0,1\}$ is a treatment assignment indicator that is 1 if unit i was assigned the treatment and is zero otherwise. C_i may be uncertain at the time of treatment assignment and is allowed to be observed after treatment assignment.

To account for heterogeneous treatment responses and costs, we work from a potential outcomes and costs framework. For unit i and for $j \in \{0,1\}$, let $Y_{i,j}$ and $C_{i,j}$ denote the outcome and cost, respectively, that would have been observed if unit i had been assigned $D_i = j$. Ignoring the index i, we can relate the observed outcome and cost to their potential outcomes and costs by writing

$$Y = Y_1 D + Y_0 (1 - D),\ \ C = C_1 D + C_0 (1 - D). \tag{1.1}$$

The following assumption formalizes this setting. It also includes conditions needed to identify properties related to potential outcomes and costs when they are not observed directly in sample data.

Assumption 1.3.1 *(i) Random Sample: Let Q be the joint distribution of $(Y_0, Y_1, C_0, C_1, D, X)$, where $Y_0, Y_1, C_0, C_1 \in \mathbb{R}$, $D \in \{0,1\}$, $X \in \mathscr{X} \subseteq \mathbb{R}^{d_x}$. Let $Z = (Y, C, D, X) \in \mathscr{Z}$ be distributed according to P where P is determined by Q and (1.1). We assume the sample $S = \{Z_i\}_{i=1}^n \sim P^{\otimes n}$ is a size n i.i.d. sample[1]. We denote the sample space $S \in \mathscr{S} = \mathscr{Z}^n$.*

(ii) Unconfoundedness: $(Y_1, Y_0, C_1, C_0) \perp D | X$.

(iii) Bounded Outcomes and Costs: There exist positive $M_y, M_c < \infty$ such that the support of Y is contained in $[-M_y/2, M_y/2]$ and the support of C is contained in $[-M_c/2, M_c/2]$.

(iv) Strict overlap: Define $e(X) = E_P[D|X]$, where $E_P(\cdot)$ is the expectation with respect to P.[2] It is assumed that there exists $\kappa \in (0, 1/2)$ such that $e(x) \in [\kappa, 1-\kappa]$ for all $x \in \mathscr{X}$.

Assumption 1.3.1 mirrors treatment assumptions in Kitagawa and Tetenov (2018) and Mbakop and Tabord-Meehan (2021) and also includes similarly-formulated conditions for cost-related variables. Unconfoundedness states that, conditional on the covariates, the potential outcomes and costs are independent of the treatments assigned to the observed data. This and strict overlap will hold in randomized controlled trials (RCTs) which is our primary setting of interest. As such, we assume $e(x)$ is known. It is possible to adjust our procedures to a setting where $e(x)$ is estimated similarly to the e-hybrid rules utilized in Kitagawa and Tetenov (2018) and Mbakop and Tabord-Meehan (2021) while maintaining some of the theoretical motivations considered in Section 1.4. We leave a complete exploration of this topic to future research and work under the presumption that $e(x)$ is known.

[1] To denote the probability of an event A under this sampling distribution, we will use the notation $P^n(A)$. To denote the probability of an event B under the distribution P, we write $P(B)$.

[2] Similarly, we denote expectation with respect to Q by $E_Q(\cdot)$. Expectation with respect to the distribution of the sample, $P^{\otimes n}$, will be denoted $E_{P^n}(\cdot)$.

Define the conditional average treatment effect (CATE) and the conditional average treatment cost (CATC), respectively, by

$$\delta_y(x) \equiv E_Q[Y_1 - Y_0|X = x], \ \ \delta_c(x) \equiv E_Q[C_1 - C_0|X = x]. \tag{1.2}$$

Assumption 1.3.1 (iii) implies that $|\delta_y(X)|$ and $|\delta_c(X)|$ are bounded almost surely by M_y and M_c, respectively. Our procedures can be implemented without knowledge of M_y or M_c and several of the motivating regret bounds in Section 1.4 could be derived in slightly altered forms if instead we required that objects related to $|\delta_y(X)|$ and $|\delta_c(X)|$ are sub-Guassian or even sub-exponential with additional constraints on a hyper-parameter. Assumption 1.3.1 (iii) is typically a mild requirement that is often adopted in the treatment and classification literature; here it simplifies our exposition and path to generalization bounds. Note that Y and C may belong to any interval. The upper and lower bounds are taken to be symmetric around zero for convenience and without loss of generality.

In section 1.3.2 we propose treatment assignment rules that aim to balance two prevailing objectives. We seek rules that will maximize the expected outcome Y while also accounting for a potential budget constraint when we anticipate that resource, policy, or other limitations may preclude treating everyone with a positive CATE. Our proposed rules contain a parameter u, which can be chosen in a data-dependent manner, that modulates how much the second (budgetary) objective is prioritized. In particular, any choice of u corresponds to a different maximum expected cost in a budget-constrained welfare optimization problem. Before describing the treatment model and empirical approach, we first state the policy maker's problem at the population level under a given maximum budget B if the distribution Q were known.

The policy maker's goal is to obtain a treatment rule that maximizes welfare subject to a budget or quantity constraint. The treatment rule is intended for application to a target population wherein the joint distribution of (Y_0, Y_1, C_0, C_1, X) follows that associated with Q. We will consider stochastic treatment assignment rules, defining such a rule as a measurable map

$f : \mathscr{X} \to [0,1]$ from the covariate space to a treatment assignment probability. If $f(x) \in \{0,1\}$, the treatment assignment for x is non-random. If $0 < f(x) < 1$, treatment is assigned randomly with treatment probability $f(x)$.

The utilitarian welfare associated with f is given by

$$E_Q[Y_1 f(X) + Y_0(1 - f(X))]. \tag{1.3}$$

This is the expected value of Y when treatment is administered according to $f(X)$. Dropping terms that do not vary with f, the policy maker's objective function evaluated at f is defined by

$$W(f) \equiv E_Q[(Y_1 - Y_0) f(X)]. \tag{1.4}$$

Choosing f that maximizes $W(f)$ is equivalent to choosing f that maximizes utilitarian welfare. Thus we will refer to $W(f)$ as the welfare associated with f. Note that by the law of iterated expectations, $W(f) = E_Q[\delta_y(X) f(X)]$. Next, define the expected cost of f by

$$K(f) \equiv E_Q\left[(C_1 - C_0) f(X)\right], \tag{1.5}$$

which can similarly be written $K(f) = E_Q[\delta_c(X) f(X)]$. Given a budget constraint B, the policy maker's problem is to identify

$$f_B^* \in \arg\max_f \left\{W(f) : K(f) \leq B\right\}, \tag{1.6}$$

where the maximization is taken over all measurable functions from $\mathscr{X}$ to $[0,1]$.

Note that $K(f) = E_Q[C_1 f(X) + C_0(1 - f(X))] - E_Q[C_0]$. The budget constraint states that the expected additional cost due to implementing treatment policy f, that beyond what would be expected if treatment were never assigned, cannot exceed B. This is flexible, as it allows for cost savings (i.e. when $C_1 < C_0$ with positive probability) to be factored into the budget. Provided such savings are possible, a policy maker could be interested in, for example, $B = 0$. In this scenario the policy maker is looking for treatment policies that may improve welfare without increasing the expected cost beyond the setting were no treatments are administered. On the other hand, if the policy maker has a fixed budget allocated to treatments and cost savings do not feed back into the budget, one can simply define $C_0 = 0$, so that the observed C is equal to the cost of treatment when treatment is provided and is zero otherwise. If there is a a fixed quantity constraint consisting of a set number of treatments and no other budgetary concerns, one can set $C_0 = 0$ and $C_1 = 1$ so that the observed C is the treatment indicator. In this case B denotes the maximum proportion of the target population for which treatments are available.

If there is no budget constraint and the policy maker is able to choose any measurable $f : \mathscr{X} \to [0,1]$, it is straightforward to verify that an optimal treatment allocation rule is given by

$$f^*(x) = 1\{\delta_y(x) > 0\}. \tag{1.7}$$

f^* assigns treatment to any unit with a positive CATE. Here, and throughout the paper, the indicator function $1\{A\}$ takes the value 1 if event A occurs and is zero otherwise. Given a particular budget constraint B, a solution to the policy maker's problem is characterized in the following theorem.

Theorem 1.3.1 *Let (Y_0, Y_1, C_0, C_1, X) be distributed according to Q. Assume that $E_Q|\delta_y(X)| < \infty$, $E_Q|\delta_c(X)| < \infty$, and that $B > E_Q[\delta_c(X)1\{\delta_c(X) < 0\}]$. Then there exist constants $\eta_B \geq 0$ and $a_1, a_2 \in [0,1]$ such that*

$$f_B^*(x) = \begin{cases} 0 & \text{if } \delta_y(x) < \eta_B \delta_c(x), \\ a_1 1\{\delta_c(x) > 0\} + a_2 1\{\delta_c(x) < 0\} & \text{if } \delta_y(x) = \eta_B \delta_c(x), \\ 1 & \text{if } \delta_y(x) > \eta_B \delta_c(x), \end{cases} \tag{1.8}$$

satisfies (1.6). *In particular, if $K(f^*) \leq B$, then one can take $\eta_B = a_1 = a_2 = 0$ and $f_B^* = f^*$; if*

$K(f^*) > B$ then (η_B, a_1, a_2) are chosen such that $K(f_B^*) = B$. If $E_Q[1\{\delta_y(X) = \eta_B \delta_c(X)\}] = 0$, f_B^* is deterministic and is the unique budget-constrained, welfare-optimizing policy in the sense that for any f' satisfying (1.6) *it holds that $f'(X) = f_B^*(X)$ a.s.*

The choice of η_B in Theorem 1.3.1 is unique, however in general there may be different choices of a_1, a_2 that produce optimal rules when $E_Q[1\{\delta_y(X) = \eta_B \delta_c(X)\}] \neq 0$. Apart from this difference, Theorem 1.3.1 is a generalization of a result in Sun et al. (2021) which restricts itself to the setting where $C_1 \geq C_0$ almost surely. In practice, of course, Q is unknown to the researcher who must estimate a suitable model f empirically. Section 1.3.2 introduces the PAC-Bayesian setting for the empirical strategy we employ.

When $E_Q[1\{\delta_y(X) = \eta_B \delta_c(X)\}] = 0$, for example when $\delta_y(X)$ and $\delta_c(X)$ have bounded densities, Theorem 1.3.1 says the optimal treatment rule is deterministic and unique in terms of the resulting treatment decisions. However, the function $\delta_y(x) - \eta_B \delta_c(x)$ in the optimal rule in this setting, given by

$$f_B^*(x) = 1\{\delta_y(x) - \eta_B \delta_c(x) > 0\},$$

is not unique. Any measurable function $m(x) : \mathscr{X} \to \mathbb{R}$ that satisfies

$$\operatorname{sign}[m(x)] = \operatorname{sign}[\delta_y(x) - \eta_B \delta_c(x)],$$

yields an optimal treatment rule via $f_m(x) = 1\{m(x) > 0\}$. This situation is similar to that in the binary forecasting problem (cf. Elliott and Lieli (2013)) and is illustrated in Figure 1.1.

In Section 1.3.2, we propose treatment rules that aggregate over a user-specified family of treatment rules in a way that is weighted towards models with high empirical budget-penalized welfare. There, we introduce Gibbs treatment rules, which aggregate over the rule family to derive a treatment probability, and related majority vote rules which aggregate over the rule family to assign treatment directly. Aside from the desirable theoretical properties derived in Section 1.4, some intuition behind such an approach is as follows. Two functions $\hat{m}(x)$ and

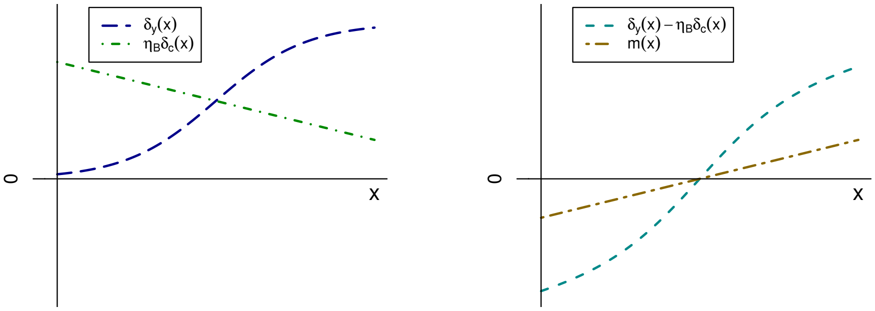


Figure 1.1. On the left, a plot of $\delta_y(x)$ and $\eta_B\delta_c(x)$ in a simple setting with a single crossing point and a single covariate. On the right, the corresponding $\delta_y(x) - \eta_B\delta_c(x)$ is plotted along with a second function, $m(x)$. Here, $m(x)$ differs from $\delta_y(x) - \eta_B\delta_c(x)$ everywhere except at the crossing point yet $1\{m(x) > 0\}$ and $1\{\delta_y(x) - \eta_B\delta_c(x)\} = f_B^*(x)$ yield identical treatment decisions.

$\hat{m}^*(x)$, with corresponding treatment rules $1\{\hat{m}(x) > 0\}$ and $1\{\hat{m}^*(x) > 0\}$, respectively, could yield identical or very similar treatment decisions over the sample covariate values. In a setting where different rules may have the same or very similar observable properties, it is reasonable to aggregate or average over rules with high empirical welfare. Rather than trying to select a single solution, we take the identification issue above as motivation for an ensemble approach.

1.3.2 Empirical Approach and PAC-Bayesian Setting

Underpinning the treatment rules we will consider is a family of non-stochastic treatment rules, indexed by $\theta \in \Theta$, denoted

$$\mathscr{F}_\Theta = \{f_\theta(x) : \mathscr{X} \to \{0,1\}; \theta \in \Theta\}. \tag{1.9}$$

For a concrete example, we could let $\{\phi_1(x), \dots, \phi_q(x)\}$ be a set of feature transformations where $\phi_j(x) : \mathscr{X} \to \mathbb{R}$ for $j = 1, \dots, q$. Denoting $\phi(x) = (\phi_1(x), \dots, \phi_q(x))^\intercal$, we could then have

$$f_\theta(x) = 1\{\phi(x)^\intercal\theta > 0\} \text{ for } \theta \in \Theta = \mathbb{R}^q, \tag{1.10}$$

where $q \in \mathbb{N}$ need not be equal to d_x, the dimension of $\mathscr{X}$.

For any treatment assignment rule f, we define the welfare regret relative to the first-best prediction rule f^* in (1.7) by

$$R(f) \equiv W(f^*) - W(f).$$

Note that $R(f)$ is defined relative to the first-best treatment assignment without a budget constraint. We can also define

$$R_B(f) \equiv W(f_B^*) - W(f), \tag{1.11}$$

the welfare-regret under a maximum expected budget of B where f_B^* is defined in Theorem 1.3.1. With simple manipulations, the oracle-type inequalities involving $R(f)$ in Sections 1.4.1 and 1.4.2 apply to $R_B(f)$ rather than $R(f)$. For simplicity, we will mostly work with $R(f)$ which is non-negative. Note that $R_B(f)$ is only non-negative when attention is constrained to treatment rules with a maximal budget B. For particular models $f_\theta \in \mathscr{F}_\Theta$, with a slight abuse of notation, we will write

$$R(\theta) \equiv R(f_\theta), \; W(\theta) \equiv W(f_\theta), \text{ and } K(\theta) \equiv K(f_\theta).$$

Under the unconfoundedness and strict overlap conditions of Assumption 1.3.1, it holds that

$$W(f) = E_Q[(Y_1 - Y_0)f(X)] = E_P\left[\left(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\right)f(X)\right].$$

A similar statement can be written for $K(f)$, now with C in place of Y. Defining

$$\delta_{y,i} = \left(\frac{Y_iD_i}{e(X_i)} - \frac{Y_i(1-D_i)}{1-e(X_i)}\right) \text{ and } \delta_{c,i} = \left(\frac{C_iD_i}{e(X_i)} - \frac{C_i(1-D_i)}{1-e(X_i)}\right),$$

the (unbiased) empirical counterparts of $W(f)$, $R(f)$, and $K(f)$, along with their notation for

$f_\theta \in \mathscr{F}_\Theta$, are given by

$$\begin{aligned}
W_n(f) &\equiv \frac{1}{n}\sum_{i=1}^{n}\delta_{y,i} f\left(X_i\right), & W_n(\theta) &\equiv \frac{1}{n}\sum_{i=1}^{n}\delta_{y,i} f_\theta\left(X_i\right), \\
R_n(f) &\equiv \frac{1}{n}\sum_{i=1}^{n}\delta_{y,i}\left(f^*\left(X_i\right) - f\left(X_i\right)\right), & R_n(\theta) &\equiv \frac{1}{n}\sum_{i=1}^{n}\delta_{y,i}\left(f^*\left(X_i\right) - f_\theta\left(X_i\right)\right), \\
K_n(f) &\equiv \frac{1}{n}\sum_{i=1}^{n}\delta_{c,i} f(X_i), & K_n(\theta) &\equiv \frac{1}{n}\sum_{i=1}^{n}\delta_{c,i} f_\theta(X_i).
\end{aligned}$$

As f^* is unknown, the empirical regret $R_n(f) = W_n(f^*) - W_n(f)$ or $R_n(\theta)$ for $\theta \in \Theta$ cannot be evaluated in practice. $R_n(\theta)$ will arise in our analysis only as a theoretical object in relation to $R(\theta)$. We stress that the treatment assignment rules we consider can be expressed solely in terms of $W_n(\theta)$.

$\mathscr{F}_\Theta$ consisting of treatment rules of the form in (1.10) will be considered in Sections 1.4.2 and 1.5. In general, to accommodate broader treatment rule model families, we make the following technical assumptions.

Assumption 1.3.2 *(i) We assume that $(\Theta, \mathscr{B}_\theta)$ is a standard Borel space. (ii) We assume that $\mathscr{F}_\Theta$ is such that the maps $(S, \theta) \mapsto R_n(\theta) : \mathscr{S} \times \Theta \to \mathbb{R}$ and $(S, \theta) \mapsto K_n(\theta) : \mathscr{S} \times \Theta \to \mathbb{R}$ are measurable.*

We now introduce the stochastic treatment rules of interest. Let $\mathscr{P}(\Theta)$ be the set of probability measures on $(\Theta, \mathscr{B}_\theta)$ and, for any $\pi \in \mathscr{P}(\Theta)$, let $\mathscr{P}_\pi(\Theta) = \{\rho \in \mathscr{P}(\Theta) : \rho \ll \pi\}$. That is, $\mathscr{P}_\pi(\Theta)$ is the set of probability measures on $(\Theta, \mathscr{B}_\theta)$ that are absolutely continuous with respect to π. Rather than selecting a single value $\hat{\theta} \in \Theta$, for example that which maximizes $W_n(\theta)$, and then assigning treatment via $f_{\hat{\theta}}$, we seek probability measures $\rho \in \mathscr{P}(\Theta)$ from which we form stochastic treatment rules. Borrowing nomenclature from the classification literature, we work with Gibbs treatment rules. For $\rho \in \mathscr{P}(\Theta)$, the Gibbs treatment rule or method associated ρ, denoted $f_{G,\rho} : \mathscr{X} \to [0,1]$, is defined by

$$f_{G,\rho}(x) = \int_\Theta f_\theta(x) d\rho(\theta),\ x \in \mathscr{X}.$$

Assigning treatments via the Gibbs method is equivalent to assigning treatments as follows. For an individual with covariates X, a parameter value $\theta_\circ$ is drawn randomly according to ρ, i.e. $\theta_\circ \sim \rho$. Then, $f_{\theta_\circ}(X) \in \{0,1\}$ determines the treatment assignment. This process, with an independent draw from ρ, is repeated each time treatment is to be assigned. Note that, exchanging the order of integration, we can write

$$R(f_{G,\rho}) = \int_\Theta R(\theta)d\rho(\theta) \text{ and } R_n(f_{G,\rho}) = \int_\Theta R_n(\theta)d\rho(\theta),$$

which is called the Gibbs risk associated with ρ. Similarly, the expected cost of $f_{G,\rho}$ and its empirical counterpart can be written

$$K(f_{G,\rho}) = \int_\Theta K(\theta)d\rho(\theta) \text{ and } K_n(f_{G,\rho}) = \int_\Theta K_n(\theta)d\rho(\theta).$$

We will frequently be concerned with the cost or empirical cost associated with a Gibbs treatment rule utilizing some $\rho \in \mathscr{P}_\pi(\Theta)$. To simplify the exposition, we denote

$$B(\rho) \equiv K\left(f_{G,\rho}\right), \text{ and } \widehat{B}(\rho) \equiv K_n\left(f_{G,\rho}\right). \tag{1.12}$$

A non-stochastic treatment rule that is closely related to the Gibbs rule is the so-called majority vote or Bayes method associated with $\rho \in \mathscr{P}(\Theta)$. This is given by

$$f_{\text{mv},\rho}(x) = 1\left\{\int_\Theta f_\theta(x)d\rho(\theta) > \frac{1}{2}\right\},\ x \in \mathscr{X}. \tag{1.13}$$

In practice, majority vote rules can deliver treatment rules that are numerically more stable than their Gibbs counterpart. If $\rho = \alpha\rho_1 + (1-\alpha)\rho_2$ for some $\rho_1, \rho_2 \in \mathscr{P}(\Theta)$ and constant α, then $R(f_{G,\rho}) = \alpha R(f_{G,\rho_1}) + (1-\alpha)R(f_{G,\rho_2})$. That is, the Gibbs risk is a linear functional of ρ. This linearity makes the Gibbs risk and Gibbs treatment rules more amenable to theoretical analysis. Our analysis will therefore focus on a family of Gibbs treatment rules. However, in Section 1.4.3, we show that the majority vote treatment rule associated with our Gibbs rules of interest inherit desirable properties from their Gibbs counterparts. In practice, either method is an acceptable choice and we consider both in our simulation study in Section 1.5.

In particular, we propose to use Gibbs treatment rules utilizing data-dependent[3] probability measures of the form $\hat{\rho}_{\lambda,u}$ defined below.

Definition 1.3.2 *For $\lambda > 0$, $u \geq 0$, and a reference measure $\pi \in \mathscr{P}(\Theta)$, define $\hat{\rho}_{\lambda,u}$ to be the (random) probability measure on Θ with the following Radon-Nikodym (RN) derivative with respect to π:*

$$\begin{aligned}\frac{d\hat{\rho}_{\lambda,u}}{d\pi}(\theta) &= \frac{\exp\left[-\lambda\left(R_n(\theta)+uK_n(\theta)\right)\right]}{\int_\Theta \exp\left[-\lambda\left(R_n\left(\tilde{\theta}\right)+uK_n\left(\tilde{\theta}\right)\right)\right]d\pi\left(\tilde{\theta}\right)} \\ &= \frac{\exp\left[-\lambda\left(uK_n(\theta)-W_n(\theta)\right)\right]}{\int_\Theta \exp\left[-\lambda\left(uK_n\left(\tilde{\theta}\right)-W_n\left(\tilde{\theta}\right)\right)\right]d\pi\left(\tilde{\theta}\right)}.\end{aligned}$$

Define $\rho^*_{\lambda,u}$ to be the probability measure on Θ with the following RN derivative with respect to π:

$$\frac{d\rho^*_{\lambda,u}}{d\pi}(\theta) = \frac{\exp\left[-\lambda\left(R(\theta)+uK(\theta)\right)\right]}{\int_\Theta \exp\left[-\lambda\left(R\left(\tilde{\theta}\right)+uK\left(\tilde{\theta}\right)\right)\right]d\pi\left(\tilde{\theta}\right)}.$$

$\hat{\rho}_{\lambda,u}$ is sometimes called a Gibbs posterior distribution or a Boltzmann distribution. As $\lambda \to \infty$, $\hat{\rho}_{\lambda,u}$ concentrates around the value of θ such that f_θ minimizes the budget-penalized empirical regret criterion $R_n(f_\theta)+uK_n(f_\theta)$. Equivalently, it concentrates around the value of θ the maximizes $W_n(\theta)-uK_n(\theta)$ over Θ. This reduces to the empirical welfare maximizer when $u=0$. In general, $\hat{\rho}_{\lambda,u}$ assigns higher probability to regions of the parameter or model space with low budget-penalized empirical regret. u modulates the trade off between emphasis on low regret vs expected cost. As subsequent analysis will show, different choices of u correspond

[3] In general, by data-dependent probability measures on $(\Theta,\mathscr{B}_\theta)$ we mean regular conditional probability measures (RCPMs): letting $\mathscr{B}_s$ denote the σ-algebra associated with the sample space $\mathscr{S}$, $\rho(S,\cdot)$ is an RCPM on $(\Theta,\mathscr{B}_\theta)$ if (i) for any fixed $A \in \mathscr{B}_\theta$, the map $S \mapsto \rho(S,A) : (\mathscr{S},\mathscr{B}_s) \to \mathbb{R}_+$ is measurable; and (ii) for any $S \in \mathscr{S}$, the map $A \mapsto \rho(S,A) : \mathscr{B}_\theta \to [0,1]$ is a probability measure. For additional measure-theoretic details, for example the decomposition and measurability of the Kullback-Leibler divergence (utilized throughout the paper) between RCPMs, we refer the reader to Catoni (2004), in particular Proposition 1.7.1 and its proof on pages 50-54.

in a one-to-one manner with different budget constraints. We will consider the setting where u is cross-validated and the setting where it is determined by a particular choice of a budget constraint parameter B. $\rho^*_{\lambda,u}$ is a theoretical counterpart to $\hat{\rho}_{\lambda,u}$ that will be useful when we analyze statistical properties related to $\hat{\rho}_{\lambda,u}$. λ is typically chosen via cross-validation while choices where $\lambda = \mathscr{O}(\sqrt{n})$ will yield optimal or near-optimal rates of convergence in Section 1.4.

In the PAC-Bayesian literature, probability measures over the model or parameter space that are traditionally chosen independently of the sample are often called prior probability measures. In our setting, the choice of π utilized in Definition 1.3.2 will fall into this category. Probability measures utilized for treatment or prediction, such as $\hat{\rho}_{\lambda,u}$, are called posterior distributions. However, this nomenclature does not have the same connotation as in traditional Bayesian methodology. While knowledge of the DGP could allow for a prior to be chosen that improves the performance of rules suggested from PAC-Bayesian analysis, often the prior is taken to be uniform or normal centered at the origin. Additionally, the posterior, for example, does not need to be proportional to a likeilihood function. The statistical analysis itself is frequentist in nature. The role and choice of π will be discussed further later in the paper. For now we make the following assumption.

Assumption 1.3.3 $\pi \in \mathscr{P}(\Theta)$ *is a (deterministic) probability measure that does not depend on the sample.*

1.3.3 Initial Properties of the Gibbs Posterior

Here we derive initial properties of $\hat{\rho}_{\lambda,u}$ that link the choice of u to a particular budget constraint. These provide intuition behind Definition 1.3.2 and are utilized in proving the results of Section 1.4.

Let $D_{\mathrm{KL}}(\rho,\pi)$ denote the Kullback–Leibler (KL) divergence between $\rho,\pi \in \mathscr{P}(\Theta)$,

$$D_{\mathrm{KL}}(\rho,\pi) = \begin{cases} \int_\Theta \log\left[\frac{d\rho}{d\pi}(\theta)\right] d\rho(\theta), & \text{if } \rho \ll \pi \\ \infty, & \text{else.} \end{cases}$$

Suppose the policy maker has a maximum expected budget of $B \in \mathbb{R} \cup \{\infty\}$, where $B = \infty$ is the unconstrained setting. If the data generating process were known, among Gibbs treatment rules we would be interested in a solution to

$$\min_{\rho \in \mathscr{P}(\Theta)} \int_\Theta R(\theta) d\rho(\theta), \text{ subject to } \int_\Theta K(\theta) d\rho(\theta) \leq B. \tag{1.14}$$

In practice, we will instead focus on a subset $\mathscr{P}_\pi(\Theta) \subset \mathscr{P}(\Theta)$ and solve the following empirical problem:

$$\min_{\rho \in \mathscr{P}_\pi(\Theta)} \left[\int_\Theta R_n(\theta) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho,\pi)\right], \text{ subject to } \int_\Theta K_n(\theta) d\rho(\theta) \leq B. \tag{1.15}$$

(1.15) includes a regularization term in the form of $D_{\mathrm{KL}}(\rho,\pi)$, discouraging any choice for ρ that has a large KL divergence from the reference measure π. In practice, $\mathscr{P}_\pi(\Theta)$ is flexible and optimal choices for λ will entail $\lambda \to \infty$ as $n \to \infty$. When adapted to our setting, Lemma 1.3.1 below shows that, provided a feasibility or Slater condition holds, for some value $\hat{u} \geq 0$, $\hat{\rho}_{\lambda,\hat{u}}$ is the solution to (1.15). Of course, appearing to be a reasonable empirical counterpart of (1.14) is not, in and of itself, justification for $f_{G,\hat{\rho}_{\lambda,u}}$. In Section 1.4 we provide additional theoretical motivation for $f_{G,\hat{\rho}_{\lambda,u}}$, comparing it to alternative Gibbs rules and optimal (non-stochastic) models in $\mathscr{F}_\Theta$.

The following lemma yields solutions to (1.15) and a theoretical counterpart when $R_n(\theta)$ and $K_n(\theta)$ are replaced by $R(\theta)$ and $K(\theta)$, respectively.

Lemma 1.3.1 *Let $\pi \in \mathscr{P}(\Theta)$, $\lambda > 0$, $B \in \mathbb{R} \cup \{\infty\}$, and let $A(\theta)$ and $H(\theta)$ be bounded, mea-*

surable functions defined on $(\Theta, \mathscr{B}_\Theta)$*. For* $u \geq 0$*, define* $\tilde{\rho}_{A,H,\lambda,u} \in \mathscr{P}_\pi(\Theta)$ *to be the probability measure with RN derivative with respect to* π *given by*

$$\frac{d\tilde{\rho}_{A,H,\lambda,u}}{d\pi}(\theta) = \frac{\exp\left[-\lambda\left(A(\theta) + uH(\theta)\right)\right]}{\int_\Theta \exp\left[-\lambda\left(A\left(\tilde{\theta}\right) + uH\left(\tilde{\theta}\right)\right)\right] d\pi\left(\tilde{\theta}\right)}.$$

Lastly, define

$$\Lambda(u) = \int_\Theta H(\theta) d\tilde{\rho}_{A,H,\lambda,u}(\theta),\ u \geq 0,\ \text{ and }\ \mathscr{E}_{H,B} = \left\{\rho \in \mathscr{P}_\pi(\Theta) : \int_\Theta H(\theta) d\rho(\theta) \leq B\right\}.$$

We have the following result. If

$$\pi\left(\{\theta : H(\theta) < B\}\right) > 0, \tag{1.16}$$

then,

$$\tilde{\rho}_{A,H,\lambda,\overline{u}_B} = \underset{\mathscr{E}_{H,B}}{\arg\min}\left[\int_\Theta A(\theta)\, d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi)\right], \tag{1.17}$$

where $\overline{u}_B = 0$ *if* $\Lambda(0) \leq B$ *and otherwise, when* $\Lambda(0) > B$*,* $\overline{u}_B > 0$ *is the unique positive real number satisfying* $\Lambda(\overline{u}_B) = B$*. Additionally*[4]*,*

$$\begin{aligned} &\int_\Theta A(\theta) d\tilde{\rho}_{A,H,\lambda,\overline{u}_B}(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}\left(\tilde{\rho}_{A,H,\lambda,\overline{u}_B}, \pi\right) \\ &= \sup_{u \geq 0}\left[\int_\Theta A(\theta) d\tilde{\rho}_{A,H,\lambda,u}(\theta) + u\left(\int_\Theta H(\theta) d\tilde{\rho}_{A,H,\lambda,u}(\theta) - B\right) + \frac{1}{\lambda} D_{\mathrm{KL}}\left(\tilde{\rho}_{A,H,\lambda,u}, \pi\right)\right]. \end{aligned} \tag{1.18}$$

When $B = \infty$, so that $\overline{u}_B = 0$, the result in Lemma 1.3.1 is a well known property that is commonly utilized in the PAC-Bayesian literature with $A(\theta)$ taken as some loss or regret function; see Catoni (2007) and Alquier et al. (2016) among many possible examples. Lemma 1.3.1 extends this setting to accommodate a secondary constraint objective associated with

[4] Throughout, we adopt the convention that $0 \cdot -\infty = 0$ when $B = \infty$ in statements of this form.

$H(\theta)$. When, for example $H(\theta) = R(\theta)$, $\Lambda(u)$ is the cost associated with the Gibbs treatment rule utilizing $\tilde{\rho}_{A,H,\lambda,u}$. $\Lambda(u)$ is decreasing in u. Intuitively, as the exponential re-weighting of π depends more heavily on $H(\theta)$ for larger values of u, regions of the parameter or model space with greater cost receive a relatively lower weighting and the overall cost is reduced as u increases. Convex optimization problems where the objective or constraint set involves the Kullback-Liebler divergence have been considered in earlier work, for example in Csiszár (1975). Rather than establishing Lemma 1.3.1 from the more abstract setting there, the proof in the Appendix utilizes well known properties of the KL divergence, stated as Lemma 1.A.1 and Corollary 1.A.1 in the Appendix. We note that Corollary 1.A.1 (b) is a well known change-of-measure inequality (c.f. Csiszár (1975) and Donsker and Varadhan (1975)) that is widely utilized in deriving PAC-Bayesian generalization bounds.

The property in (1.18) is used in deriving the oracle-type inequalities in Section 1.4. The result states that the duality gap between the primal and dual of the minimization problem in (1.17) is zero. That is,

$$\begin{aligned}
&\min_{\rho\in\mathscr{P}_\pi(\Theta)} \sup_{u\geq 0} \left[\int_\Theta A(\theta)d\rho(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi) + u\left(\int_\Theta H(\theta)d\rho(\theta) - B\right)\right] \\
&= \sup_{u\geq 0} \min_{\rho\in\mathscr{P}_\pi(\Theta)} \left[\int_\Theta A(\theta)d\rho(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi) + u\left(\int_\Theta H(\theta)d\rho(\theta) - B\right)\right].
\end{aligned}$$

Note that the left-hand side of the above equality, the primal problem, is equivalent to the optimization problem in (1.17). The right-hand side is the dual of this problem. That the right-hand side above is equivalent to the expression on the right-hand side of (1.18) can be seen from a careful examination of (1.17) or from Corollary 1.A.1 (a) in the Appendix. The condition in (1.16) constitutes a constraint qualification.

We will apply Lemma 1.3.1 with $A(\theta) = R_n(\theta)$ or $R(\theta)$ and $H(\theta) = K_n(\theta)$ or $K(\theta)$. We consider two scenarios or perspectives. In the first, we have a (nonrandom) predetermined budget B and utilize a corresponding, sample dependent, choice of $\hat{u}$. In the second scenario, we start from a predetermined, non-random choice of u (or multiple values of u), which then corresponds to a sample dependent budget (or budgets) associated with $f_{G,\hat{\rho}_{\lambda,u}}$. We will require the following assumptions in order to satisfy (1.16) in our analysis. The first will correspond to the case with a predetermined B while the second condition will be utilized when we start from predetermined u.

Assumption 1.3.4 *(i) Let $B \in \mathbb{R} \cup \{\infty\}$ be a desired budget. It is assumed that*

$$\pi(\theta \in \Theta : K(\theta) < B) > 0 \text{ and } \pi(\theta \in \Theta : K_n(\theta) < B) > 0 \;\; P^n \text{ almost surely.}$$

(ii) It is assumed that

$$\mathbb{V}_{\theta\sim\pi}[K(\theta)] > 0 \text{ and } \mathbb{V}_{\theta\sim\pi}[K_n(\theta)] > 0 \;\; P^n \text{ almost surely}$$

where, $\mathbb{V}_{\theta\sim\pi}$ denotes the variance of $K(\theta)$ when $\theta \sim \pi$ and, for a fixed sample $S \in \mathscr{S}$, $\mathbb{V}_{\theta\sim\pi}[K_n(\theta)]$ denotes the variance of $K_n(\theta)$ when $\theta \sim \pi$.

Assumption 1.3.4 involves $\mathscr{F}_\Theta$, π and the sampling distribution P. Condition (i) requires that the budget of interest is not ruled out under the prior or reference measure π and is not exactly at the boundary of theoretical or empirical feasibility. With additional exposition, the condition that $\pi(\theta \in \Theta : K_n(\theta) < B) > 0$ holds P^n a.s. could be replaced by the condition that $\pi(\theta \in \Theta : K_n(\theta) < B) > 0$ holds with high probability. For example, with probability at least $1-\xi$, for some $\xi \in [0,1)$. In this case the theorems in Section 1.4 will remain valid except that the high probability bounds there, that hold with probability at least $1-\varepsilon$ for $\varepsilon \in (0,1]$, will now hold with probability at least $1-\varepsilon-\xi$. Condition (ii) requires that there is always variation in actual and empirical costs within models in $\mathscr{F}_\Theta$ drawn by π.

Given Lemma 1.3.1 and the assumption above, the following definition will be relevant when the analysis starts with a predetermined budget B for which we must find an appropriate value of u.

Definition 1.3.3 *Let $\hat{\rho}_{\lambda,u}$ and $\rho^*_{\lambda,u}$ be defined with $\pi \in \mathscr{P}(\Theta)$ as in Definition 1.3.2. For $B \in \mathbb{R}$, define $\hat{u}(B,\lambda)$ by*

$$\hat{u}(B,\lambda) = \underset{u \geq 0}{\arg\max} \int_{\Theta} R_n(\theta) d\hat{\rho}_{\lambda,u}(\theta) + u \left(\int_{\Theta} K_n(\theta) d\hat{\rho}_{\lambda,u}(\theta) - B \right) + \frac{1}{\lambda} D_{\mathrm{KL}} \left(\hat{\rho}_{\lambda,u}, \pi \right),$$

$$u^*(B,\lambda) = \underset{u \geq 0}{\arg\max} \int_{\Theta} R(\theta) d\rho^*_{\lambda,u}(\theta) + u \left(\int_{\Theta} K(\theta) d\rho^*_{\lambda,u}(\theta) - B \right) + \frac{1}{\lambda} D_{\mathrm{KL}} \left(\rho^*_{\lambda,u}, \pi \right).$$

For $B = \infty$, define $\hat{u}(\infty,\lambda) = 0$ and $u^*(\infty,\lambda) = 0$.

To conclude the section, we point out corollaries of Lemma 1.3.1 and Assumption 1.3.4 relevant to our setting. Define the sets

$$\mathscr{E}_B = \left\{ \rho \in \mathscr{P}_\pi(\Theta) : \int_{\Theta} K(\theta) d\rho(\theta) \leq B \right\},\ B \in \mathbb{R} \cup \{\infty\} \tag{1.19}$$

and

$$\widehat{\mathscr{E}}_B = \left\{ \rho \in \mathscr{P}_\pi(\Theta) : \int_{\Theta} K_n(\theta) d\rho(\theta) \leq B \right\},\ B \in \mathbb{R} \cup \{\infty\}. \tag{1.20}$$

In the scenario where we start from a pre-selected B, $\mathscr{E}_B$ is the (non-random) subset of $\mathscr{P}_\pi(\Theta)$ corresponding to Gibbs treatment rules with expected cost within the budget. $\widehat{\mathscr{E}}_B$ a random set that serves as an empirical counterpart, denoting the $\rho \in \mathscr{P}_\pi(\Theta)$ with Gibbs rules that meet the budget constraint empirically.

When analysis begins with a pre-determined value of u, $B(\hat{\rho}_{\lambda,u})$ as in Assumption 1.3.4 and its empirical counterpart $\widehat{B}(\hat{\rho}_{\lambda,u})$ both defined in (1.12), are both random. $B(\hat{\rho}_{\lambda,u})$ is the expected cost of $f_{G,\hat{\rho}_{\lambda,u}}$ in the target population given the sample-dependent $\hat{\rho}_{\lambda,u}$. This is not observed. However, it is a key object of interest, as it tells the researcher the expected cost of the estimated policy $f_{G,\hat{\rho}_{\lambda,u}}$ associated with u. Similarly, for a predetermined u, both $\mathscr{E}_{B(\hat{\rho}_{\lambda,u})}$ and $\widehat{\mathscr{E}}_{\widehat{B}(\hat{\rho}_{\lambda,u})}$ are random sets. The former corresponds to all Gibbs treatment policies with an expected budget in the target population that is less than or equal to that of $f_{G,\hat{\rho}_{\lambda,u}}$. The latter serves as an empirical counterpart for which membership can be evaluated from the sample.

Given Lemma 1.3.1 and Assumption 1.3.4, the following lemma pertains to the empirical problem in (1.15) and is mostly a corollary to 1.3.1. It says that, for a pre-specified B, $\hat{\rho}_{\lambda,\hat{u}(B,\lambda)}$ solves (1.15). Conversely, if we start with a predetermined value of u, $\hat{\rho}_{\lambda,u}$ solves an analogous problem where the budget is given by $\widehat{B}(\hat{\rho}_{\lambda,u})$.

Lemma 1.3.2 *(a) Let Assumptions 1.3.2 and 1.3.4 (i) hold for $B \in \mathbb{R} \cup \{\infty\}$. The following properties hold P^n almost surely. For any $\lambda > 0$, $\hat{u}(B,\lambda)$ exists, is unique, and satisfies that $\hat{u}(B,\lambda) = 0$ when $\int_\Theta K_n(\theta) d\hat{\rho}_{\lambda,0}(\theta) \leq B$ and $\hat{u}(B,\lambda)$ is positive and satisfies $\int_\Theta K_n(\theta) d\hat{\rho}_{\lambda,\hat{u}}(\theta) = B$ when $\int_\Theta K_n(\theta) d\hat{\rho}_{\lambda,0}(\theta) > B$. Additionally,*

$$\hat{\rho}_{\lambda,\hat{u}(B,\lambda)} = \underset{\widehat{\mathscr{E}}_B}{\arg\min} \left[\int_\Theta R_n(\theta) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho,\pi) \right],$$

(b) Let Assumptions 1.3.2 and 1.3.4 (ii) hold. Then, P^n almost surely,

$$\hat{\rho}_{\lambda,u} = \underset{\widehat{\mathscr{E}}_{\widehat{B}(\hat{\rho}_{\lambda,u})}}{\arg\min} \left[\int_\Theta R_n(\theta) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho,\pi) \right].$$

1.4 PAC-Bayesian Analysis

Here we provide theoretical motivation for decision rules utilizing $\hat{\rho}_{\lambda,u}$ or $\hat{\rho}_{\lambda,\hat{u}(B,\lambda)}$. In Section 1.4.1, we first construct PAC-Bayesian generalization bounds that are similar to counterparts in earlier literature. Then we derive oracle-type inequalities that compare the proposed treatment rules to alternatives in terms of regret in the target population for a given budget. The results in Section 1.4.1 allow for a general choice of the prior or reference measure π utilized in the definition of $\hat{\rho}_{\lambda,u}$ and $\hat{\rho}_{\lambda,\hat{u}(B,\lambda)}$. As a result, several bounds there contain KL divergence terms related to the complexity of the learning problem and the model class $\mathscr{F}_\Theta$. In Section 1.4.2, we specify $\mathscr{F}_\Theta$ to consist of rules of the form in (1.10) and take π to be an uninformative multivariate normal distribution. In this setting, we obtain oracle-type inequalities that compare the regret of our proposed treatment assignment rules directly to that of the rules in

$\mathscr{F}_\Theta$ with the lowest welfare regret that are in budget. In section 1.4.3, we show that desirable properties for the majority vote rules associated with $\hat{\rho}_{\lambda,u}$ can be inherited by their majority vote counterparts.

Our analysis builds from results and techniques in the PAC-Bayesian literature that are not always stated in ways that are directly applicable to our setting. Results from earlier literature are adapted to our setting in Appendix Section 1.A.1, which also contains additional properties of interest. For the most part, proofs are included there for completeness even when the adjustments are fairly minor. This spares the reader from visiting multiple references requiring concerted adjustments at certain steps of our analysis. Proofs specific to Section 1.4 are contained in Appendix Section 1.A.3.

1.4.1 Regret Bounds and Oracle-Type Inequalities

The first step in our analysis, Theorem 1.4.1, obtains alterations of earlier PAC-Bayesian generalization bounds for the treatment assignment setting. A variant of part (a) appears in Catoni (2007) which considers classification in the 0/1-loss setting. In our setting, it can be derived as a special case of a bound appearing in Alquier et al. (2016) or via a general approach to PAC-Bayesian bounds outlined, for example, in Germain et al. (2015). We utilize the latter approach which is useful during additional steps of our analysis. The proofs of parts (b) and (c) utilize the approach of Lever et al. (2010), with part (b) being an alteration of Theorem 3 in that work.

Theorem 1.4.1 *Let $\pi \in \mathscr{P}(\Theta)$ and let Assumptions 1.3.1, 1.3.2, and 1.3.3 hold. Set*

$$\{V_n(\theta), V(\theta), M_\ell\} = \{R_n(\theta), R(\theta), M_y\} \text{ or else } \{V_n(\theta), V(\theta), M_\ell\} = \{K_n(\theta), K(\theta), M_c\}.$$

We have the following properties.

(a) Let $\varepsilon \in (0,1]$, $\lambda > 0$ and $s \in \{-1,1\}$. With probability at least $1-\varepsilon$, for all $\rho \in$

$\mathscr{P}_\pi(\Theta)$ simultaneously it holds that

$$\int_\Theta s\left[V_n(\theta)-V(\theta)\right]d\rho(\theta) \le \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)+\frac{1}{\lambda}\left[\frac{\lambda^2 M_\ell^2}{8n\kappa^2}+\log\frac{1}{\varepsilon}\right].$$

(b) Let $\lambda > 0$, $u \ge 0$, and $\varepsilon \in (0,1]$. With probability at least $1-\varepsilon$, it holds that

$$\begin{aligned}&\left(\int_\Theta V(\theta)d\hat{\rho}_{\lambda,u}(\theta)-\int_\Theta V_n(\theta)d\hat{\rho}_{\lambda,u}(\theta)\right)^2\\ &\le \frac{M_\ell^2}{2n\kappa^2}\left[\frac{\lambda\sqrt{2}\,(M_y+uM_c)}{\kappa\sqrt{n}}\sqrt{\log\left(2\sqrt{n}\right)+\log\frac{2}{\varepsilon}}+\frac{\lambda^2\,(M_y+uM_c)^2}{2n\kappa^2}+\log\left(2\sqrt{n}\right)+\log\frac{2}{\varepsilon}\right].\end{aligned}$$

(c) Let $\lambda > 0$, $u \ge 0$, and $\varepsilon \in (0,1]$. With probability at least $1-\varepsilon$, it holds that

$$\begin{aligned}&\int_\Theta V(\theta)d\hat{\rho}_{\lambda,u}(\theta)-\int_\Theta V_n(\theta)d\hat{\rho}_{\lambda,u}(\theta)\\ &\le \frac{\sqrt{2}\,(M_y+uM_c)}{\kappa\sqrt{n}}\sqrt{\log\left(2\sqrt{n}\right)+\log\frac{2}{\varepsilon}}+\frac{\lambda\,(M_y+uM_c)^2}{2n\kappa^2}+\frac{1}{\lambda}\left[\frac{\lambda^2 M_\ell^2}{8n\kappa^2}+\log\frac{2}{\varepsilon}\right].\end{aligned}$$

Theorem 1.4.1 contains high probability bounds for notions of the generalization error between the target population regret (or alternatively, expected cost) and its empirical counterpart for Gibbs treatment rules. For example, one notion of generalization error for the cost of policy $f_{G,\hat{\rho}_{\lambda,u}}$ could be the absolute difference,

$$\left|K\left(f_{G,\hat{\rho}_{\lambda,u}}\right)-K_n\left(f_{G,\hat{\rho}_{\lambda,u}}\right)\right|.$$

Suppose we take $\lambda = a\kappa\sqrt{n}/(M_y+uM_c)$ for some constant $a>0$. Then Part (b) says that with probability at least $1-\varepsilon$, this absolute difference is less than or equal to

$$\frac{M_c}{\kappa\sqrt{2n}}\left[a\sqrt{\log(4n)+2\log\frac{2}{\varepsilon}}+\frac{a^2}{2}+\log(2\sqrt{n})+\log\frac{2}{\varepsilon}\right]^{1/2}=\mathscr{O}\left(\frac{\log n}{\sqrt{n}}\right).$$

When M_c and M_y are known, this upper bound can be evaluated for a given choice of a.

We say that $K_n(f_{G,\hat{\rho}_{\lambda,u}})$ is Probably (with probability at least $1-\varepsilon$) and Approximately (the $\mathscr{O}(\sqrt{\log(n)/n})$ upper bound on the absolute difference) Correct for $K(f_{G,\hat{\rho}_{\lambda,u}})$. This suggests that for a predetermined choice of u, $K_n(f_{G,\hat{\rho}_{\lambda,u}})$ will give a reasonable estimate of the expected cost in the target population, $K(f_{G,\hat{\rho}_{\lambda,u}})$, provided that λ is not too large. Part (c) is a variation of the style of bound in (b) that is useful in deriving subsequent results. We note that the above choice for λ may not be best in practice, or even feasible if the upper bound M_c is not known. In practice λ is chosen via cross-validation, which can be accommodated by Theorem 1.4.1 similarly to the choice of u as discussed below.

The bounds in Theorem 1.4.1 can be adjusted to accommodate the setting where λ, u, or pairs (λ, u) are selected from a finite set of values $\mathscr{W}$. With $|\mathscr{W}|$ denoting the number of elements in $\mathscr{W}$, one can apply a union bound argument similar to that in the proof of part (b). The theorem is applied once for each element of $\mathscr{W}$ with size $\varepsilon/|\mathscr{W}|$ for each repetition. Then, applying the union bound argument, the bounds as stated in Theorem 1.4.1 remain valid for any element of $\mathscr{W}$ with the alteration that the term $\log\frac{1}{\varepsilon}$ in part (a) is replaced by $(\log\frac{1}{\varepsilon}+\log|\mathscr{W}|)$ and the terms $\log\frac{2}{\varepsilon}$ in parts (b) and (c) are replaced by $(\log\frac{2}{\varepsilon}+\log|\mathscr{W}|)$. For example, when $\lambda = \mathscr{O}(\sqrt{n})$, this adds a term that is $\mathscr{O}(\log|\mathscr{W}|/\sqrt{n})$ to the right hand side of the high probability bound in part (a). This observation is applicable to the remaining theorems in the paper, with minor adjustments. Therefore, it is not unreasonable to start with multiple values for u. Then one may choose u in $\hat{\rho}_{\lambda,u}$ for the final policy based on the empirical estimates of the associated budgets, $K_n(f_{G,\hat{\rho}_{\lambda,u}})$ for $u \in \mathscr{W}$, or via cross-validation.

Before comparing our suggested treatment policies to alternative choices, we discuss a final insight from Theorem 1.4.1. Part (a) yields that, with probability at least $1-\varepsilon$,

$$R(f_{G,\rho}) \leq \int_{\Theta} R_n(\theta)d\rho(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{1}{\varepsilon}\right], \tag{1.21}$$

for all $\rho \in \mathscr{P}_\pi$ simultaneously. Given a budget B such that Assumption 1.3.4 (i) holds, Lemma 1.3.2 (a) states that $\hat{\rho}_{\lambda,\hat{u}(B,\lambda)}$ produces the smallest upper bound for the target population regret in (1.21) among all $\rho \in \mathscr{P}_\pi(\Theta)$ such that $K_n(f_{G,\rho}) \leq B$. Similarly, starting from a given value of u, under Assumption 1.3.4 (ii), Lemma 1.3.2 (b) shows that $\hat{\rho}_{\lambda,u}$ results in the smallest upper bound for the target population regret among Gibbs rules with an empirical budget less than or equal to $\widehat{B}(\hat{\rho}_{\lambda,u})$.

Although Theorem 1.4.1 (a) is most useful for our subsequent analysis, in the PAC-Bayesian literature there are alternative generalization bounds to (1.21) that apply for all $\rho \in \mathscr{P}_\pi(\Theta)$ and could be adapted to our setting. Most notably, variants of the bounds in Seeger (2002) and Catoni (2007) are fairly ubiquitous in the literature. Either directly or via a slight relaxation, these bounds also suggest choosing ρ to minimize

$$\int_\Theta R_n(\theta)d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho,\pi), \tag{1.22}$$

for some $\lambda > 0$. Hence, if we impose an empirical budget constraint these would again lead back to $\hat{\rho}_{\lambda,\hat{u}(B,\lambda)}$ and $\hat{\rho}_{\lambda,u}$. We note that Seeger's bound is utilized in our analysis to derive parts (b) and (c) of Theorem 1.4.1 and appears as Theorem 1.A.2 in Appendix Section 1.A.1. While this bound does not yield a closed form solution $\tilde{\rho}$ that minimizes an upper bound on the regret, we refer to the discussion in Thiemann et al. (2017) regarding a relaxation that suggests minimizing (1.22) with λ replaced by λn, which will yield the an equivalent minimization problem when λ is cross-validated. The style of bound in Catoni (2007), in particular Theorem 1.2.6 there, can be adapted to our setting via the approach in Germain et al. (2015) and again suggests choosing ρ to minimize (1.22).

Next we derive oracle-type inequalities that compare the target population regret associated with $\hat{\rho}_{\lambda,u}$ or $\hat{\rho}_{\lambda,\hat{u}(B,\lambda)}$ to that of alternative choices of ρ among Gibbs treatment rules within a relevant budget. It may be helpful to recall the definitions of $\mathscr{E}_B$ and $\mathscr{E}_{B(\hat{\rho}_{\lambda,u})}$ from (1.19) and (1.12),

$$\mathscr{E}_B = \left\{\rho \in \mathscr{P}_\pi(\Theta) : K\left(f_{G,\rho}\right) \leq B\right\} \text{ and } \mathscr{E}_{B(\hat{\rho}_{\lambda,u})} = \left\{\rho \in \mathscr{P}_\pi(\Theta) : K\left(f_{G,\rho}\right) \leq K\left(f_{G,\hat{\rho}_{\lambda,u}}\right)\right\}.$$

We have the following result.

Theorem 1.4.2 *Let $\pi \in \mathscr{P}(\Theta)$, $\lambda > 0$, and $\varepsilon \in (0,1]$. Under Assumptions 1.3.1, 1.3.2, and 1.3.3, we have the following properties.*

(a) Let $B \in \mathbb{R} \cup \{\infty\}$, denote $\hat{u} = \hat{u}(B,\lambda)$ and let Assumption 1.3.4 (i) hold. With probability at least $1-\varepsilon$, it holds that

$$R\left(f_{G,\hat{\rho}_{\lambda,\hat{u}}}\right) \leq \min_{\rho \in \mathscr{E}_B} \left\{ R\left(f_{G,\rho}\right) + \frac{2}{\lambda} D_{\mathrm{KL}}(\rho,\pi) \right\} + \frac{2}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right] + \hat{u}\sqrt{\frac{M_c^2 \log\frac{3}{\varepsilon}}{2n\kappa^2}}.$$

(b) Fix $u \geq 0$ and let Assumption 1.3.4 (ii) hold. With probability at least $1-\varepsilon$, it holds that

$$R\left(f_{G,\hat{\rho}_{\lambda,u}}\right) \leq \min_{\rho \in \mathscr{E}_{B(\hat{\rho}_{\lambda,u})}} \left\{ R\left(f_{G,\rho}\right) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho,\pi) \right\} + uU_1\left(\varepsilon;\lambda,u,n\right) + U_2\left(\varepsilon;\lambda,u,n\right).$$

where

$$U_1\left(\varepsilon;\lambda,u,n\right) = \frac{\sqrt{2}\left(M_y + uM_c\right)}{\kappa\sqrt{n}}\sqrt{\log\left(2\sqrt{n}\right) + \log\frac{4}{\varepsilon}} + \frac{\lambda\left(M_y + uM_c\right)^2}{2n\kappa^2},$$

and

$$U_2\left(\varepsilon;\lambda,u,n\right) = \sqrt{\frac{(M_y + uM_c)^2 \log(4/\varepsilon)}{2n\kappa^2}} + \frac{1}{\lambda}\left[\frac{\lambda^2\left(M_y^2 + uM_c^2\right)}{8n\kappa^2} + (1+u)\log\frac{4}{\varepsilon}\right].$$

Note that if $\lambda = \mathscr{O}(n^{1/2})$, then for any $u \geq 0$ and $\varepsilon \in (0,1]$,

$$U_1\left(\varepsilon;\lambda,u,n\right) = \mathscr{O}\left(\sqrt{\frac{\log(n)}{n}}\right) \text{ and } U_2\left(\varepsilon;\lambda,u,n\right) = \mathscr{O}\left(\frac{1}{\sqrt{n}}\right).$$

Theorem 1.4.2 contains sharp oracle-type inequalities that hold with high probability. They differ slightly from traditional oracle inequalities in that the right-hand sides contain objects that are random.

Consider part (b) first. In this case, the randomness on the right-hand side of the inequality stems from $\mathscr{E}_{B(\hat{\rho}_{\lambda,u})}$ which depends on the sample through $B(\hat{\rho}_{\lambda,u}) = K(f_{G,\hat{\rho}_{\lambda,u}})$, the un-observable expected target population cost of $\hat{\rho}_{\lambda,u}$. For a predetermined u, it is natural to ask if there are alternatives in $\mathscr{P}_\pi(\Theta)$ that would yield lower regret for the same or lower expected cost. $\mathscr{E}_{B(\hat{\rho}_{\lambda,u})}$ is therefore the natural set of interest for comparison with $\hat{\rho}_{\lambda,u}$ as it is the subset of $\mathscr{P}_\pi(\Theta)$ with Gibbs rules that have target population costs no greater than $B(\hat{\rho}_{\lambda,u})$. Given a budget $B(\hat{\rho}_{\lambda,u})$, an oracle with knowledge of $R(\theta)$ could solve for $\arg\min_{\rho\in\mathscr{E}_{B(\hat{\rho}_{\lambda,u})}} R(f_{G,\rho})$. For $\lambda \to \infty$, we may consider $\arg\min_{\rho\in\mathscr{E}_{B(\hat{\rho}_{\lambda,u})}} R(f_{G,\rho}) + \lambda^{-1} D_{\mathrm{KL}}(\rho,\pi)$ as a second-best oracle solution. When $\lambda = \mathscr{O}(n^{1/2})$, for example, part (b) indicates that with high probability $\hat{\rho}_{\lambda,u}$ is close to the second best oracle solution. In Section 1.4.2 we consider oracle-type inequalities without the KL penalty term appearing.

In part (a), the interpretation is similar to that in part (b), except that now the set of alternative Gibbs estimators for comparison are those that satisfy the predetermined budget B. This set is non-random, however now the right-hand side contains a term involving the random $\hat{u} = \hat{u}(B,\lambda)$. Note that $\hat{u}$ is the value taken by the Lagrange multiplier u in the problem

$$\min_{\rho\in\mathscr{E}_B} \sup_{u\geq 0} \left\{ \int_\Theta R_n(\theta) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho,\pi) + u \left(\int_\Theta K_n(\theta) d\rho(\theta) - B \right) \right\}.$$

It measures the marginal decrease in empirical penalized regret (alternatively, the increase in empirical penalized welfare) resulting from a marginal relaxation of the budget. Recall the welfare and budget are measured per treatment. For example, when benefits and costs are measured in dollars, how many dollars of penalized welfare are obtained (empirically) by increasing the maximum empirical cost by a dollar. In more extreme scenarios where a small increase in the budget produces a large increase in empirical welfare, the bound becomes less meaningful as the right-hand side approaches the maximum possible regret (if this level is exceeded, the bound becomes trivial). An example of an extreme setting would be when $\hat{u} = \mathscr{O}_p(n^\alpha)$ for some $\alpha \geq 1/2$. When $\hat{u}n^{-1/2}$ is large relative to typical or maximal values of the regret (which ranges from zero to twice the maximal welfare), this situation is visible to the analyst. For a fixed λ, a statement similar to part (a) can be obtained where $\hat{u}$ is replaced by a non-random constant if we make additional assumptions on the data generating distribution P. For example, if we instead assume the marginal increase in population penalized regret associated with a small relaxation of the empirical budget is $\mathscr{O}_p(1)$. As it stands, the bound produces a robustness check for the method's motivation. Intuitively, if it is easy to dramatically change the empirical welfare by relatively small budget changes, so that $\hat{u}n^{-1/2}$ is large, we may be in a situation where it is difficult to learn policies well for the given B and the proposed rules should be treated cautiously.

If regions of the model space with desirable regret and budget are assigned lower probability by π, the distributions $\rho \in \mathscr{P}_\pi(\Theta)$ with the best trade-off between $R(f_{G,\rho})$ and $D_{\mathrm{KL}}(\rho,\pi)$ in Theorem 1.4.2 will tend to have larger $D_{\mathrm{KL}}(\rho,\pi)$ terms. As a result, the upper bounds will be larger and less informative. Similarly, applying Theorem 1.4.1 part (a) with $\rho = \hat{\rho}_{\lambda,c}$ for either $c = u \geq 0$ or $c = \hat{u}(B,\lambda)$, and noting Lemma 1.3.2, the regret and budget bounds there are influenced by the trade-off between empirical regret (or cost) and $D_{\mathrm{KL}}(\hat{\rho}_{\lambda,c},\pi)$. $D_{\mathrm{KL}}(\hat{\rho}_{\lambda,c},\pi)$ increases when $\hat{\rho}_{\lambda,c}$ involves a greater re-weighting of π in definition 1.3.2. The impact of the KL terms in the bounds of this subsection are therefore related to the learning problem and model space complexity. It is influenced by how large the model space is, how narrow the subset of the model space with low regret/budget is, the relative difference in between lower and higher regret regions and the noisiness of the data. In parts (b) and (c) of Theorem 1.4.1, where the KL term is absent, this role falls more to the λ parameter: if the problem is more complex, larger (relative to n) values of λ are needed to achieve lower regret or cost. If λ is too large, remainder terms in the generalization error bounds increase. See Lever et al. (2010) for further discussion of complexity in the setting of bounds of the form in (b) and (c).

Conversely, when the policy maker has (sample independent) knowledge of the data generating process, they may be able to select or alter a given choice of π to focus on the regions of the model space that best balance regret and cost. Then $D_{\mathrm{KL}}(\rho,\pi)$ can be smaller for ρ that put the greatest weight on the most desirable regions of the parameter space. The result is smaller upper bounds in Theorem 1.4.2 and Theorem 1.4.1 (a). A benefit of the Gibbs rules associated with $\hat{\rho}_{\lambda,u}$ and $\hat{\rho}_{\lambda,\hat{u}(B,\lambda)}$ is that economic theory or situation-specific knowledge can be factored into the treatment rule via π. Compatibility with expert knowledge may be a valuable advantage in settings where resource limitations imply that some individuals with a positive CATE will not be treated. As we will see in Section 1.4.2, such knowledge is not required for the procedures to have desirable properties.

1.4.2 Normal Prior

As noted at the end of Section 1.4.1, perhaps unsurprisingly, knowledge about the data generating process can confer estimation benefits through the choice of π. While it is a positive attribute that the proposed treatment rules can utilize this information when available, it is important to emphasize that such knowledge is not a requirement. Learning procedures based on PAC-Bayesian analysis often utilize uninformative or less informative choices for π, such as normal distributions, uniform distributions when Θ is compatible, or sparsity inducing distributions.

Here we take π to be a multivariate normal distribution centered at the origin and utilize the models of the form in (1.10). We show that the proposed treatment rules maintain desirable properties. In doing so, the KL divergence term is removed from the oracle inequalities, resulting in a clearer comparison to alternative treatment rules. We leave an exploration of alternative prior choices and the settings where they may be desirable to future research.

We satisfy Assumptions 1.3.2 and 1.3.3 with the following, more specific, condition. Note that in the assumption below we are treating q as fixed; it does not grow with the sample size.

Assumption 1.4.1 *It is assumed that $\mathscr{F}_\Theta$ consists of treatment rules f_θ as described by* (1.10)*, with $\Theta = \mathbb{R}^q$. Let*

$$\Phi_{\mu,\sigma^2} \in \mathscr{P}(\mathbb{R}^q)$$

denote a multivariate normal distribution with mean vector μ *and covariance matrix* $\sigma^2 I_q$ *for some* $\sigma > 0$. *We assume that* $\pi = \Phi_{0,\sigma_\pi^2}$ *for some* $\sigma_\pi > 0$ *that does not depend on the sample.*

Next, we define

$$\Theta_B = \{\theta \in \mathbb{R}^q : K(\theta) \le B\} \text{ and } \Theta_{B(\hat{\rho}_{\lambda,u})} = \left\{\theta \in \mathbb{R}^q : K(\theta) \le B(\hat{\rho}_{\lambda,u})\right\},$$

and denote

$$\overline{\theta} \in \underset{\Theta_B}{\arg\min}\,[R(\theta)] \text{ and } \overline{\theta}_u \in \underset{\Theta_{B(\hat{\rho}_{\lambda,u})}}{\arg\min}\,[R(\theta)]. \tag{1.23}$$

Note that $\Theta_{B(\hat{\rho}_{\lambda,u})}$ and $\overline{\theta}_u$ are random as they vary with $B(\hat{\rho}_{\lambda,u})$. $\Theta_{B(\hat{\rho}_{\lambda,u})}$ is the set of parameters such that the corresponding models in $\mathscr{F}_\Theta$ have lower expected target population cost than $f_{G,\hat{\rho}_{\lambda,u}}$. $\overline{\theta}_u$ is the minimizer of the population regret among this set. With regard to $\overline{\theta}$ and $\overline{\theta}_u$, we assume the following condition.

Assumption 1.4.2 *With probability one,* $\overline{\theta}$ *and* $\overline{\theta}_u$ *as defined in* (1.23) *exist and are nonzero.*

This type of condition is implicitly assumed in, for example, Kitagawa and Tetenov (2018) and in Sun (2021). It simplifies the exposition rather than allowing that the models associated with these parameters have regret that is arbitrarily close to an infimum. The requirement that $\overline{\theta}$ and $\overline{\theta}_u$ are nonzero simply specifies that the covariates are relevant to the budget constrained welfare problem. Lastly, our analysis will also require the following technical condition.

Assumption 1.4.3 *There exists a constant* $\nu > 0$ *such that*

$$P\left[\left(\phi(X)^\intercal\theta\right)\left(\phi(X)^\intercal\theta'\right) < 0\right] \le \nu\|\theta - \theta'\|$$

for any θ *and* $\theta' \in \mathbb{R}^q$ *such that* $\|\theta\| = \|\theta'\| = 1$.

Assumption 1.4.3 or a direct analog is applied in several classification and bipartite ranking applications utilizing PAC-Bayesian approaches. For examples, see Ridgway et al.

(2014), Alquier et al. (2016), and Guedj and Robbiano (2018). It is a fairly mild requirement and, as is shown in Alquier et al. (2016) (c.f. p. 10 there), it is satisfied whenever $\phi(X)/\|\phi(X)\|$ has a bounded density on the unit sphere.

We have the following result.

Theorem 1.4.3 *Let Assumptions 1.3.1, 1.4.1, 1.4.2, and 1.4.3 hold. Let $\sigma_\pi = 1/\sqrt{q}$. Then we have the following properties for any $\varepsilon \in (0,1]$.*

(a) Let Assumption 1.3.4 (i) hold for a given $B \in \mathbb{R} \cup \{\infty\}$. Let $\lambda = \kappa\sqrt{nq}/M_y$, $\hat{u} = \hat{u}(B,\lambda)$ and $u^* = u^*(B,\lambda/2)$. With probability at least $1-\varepsilon$, it holds that

$$R\left(f_{G,\hat{\rho}_{\lambda,\hat{u}}}\right) \leq R\left(\overline{\theta}\right) + \sqrt{\frac{q}{n}}\log(4n)\frac{M_y}{\kappa} + \frac{2M_y\log\frac{3}{\varepsilon}}{\kappa\sqrt{nq}} + \hat{u}\sqrt{\frac{M_c^2\log\frac{3}{\varepsilon}}{2n\kappa^2}} + \frac{u^*\nu M_c}{\sqrt{n}} + \overline{U}_1(n;q),$$

where $\overline{U}_1(n;q) = \mathscr{O}(n^{-1/2})$ with the explicit formulation given in the proof.

(b) Fix $u \geq 0$ and set $\lambda = \kappa\sqrt{nq}/(M_y + uM_c)$. Let Assumption Assumption 1.3.4 (ii) hold. With probability at least $1-\varepsilon$,

$$R\left(f_{G,\hat{\rho}_{\lambda,u}}\right) \leq R\left(\overline{\theta}_u\right) + \frac{M_y + uM_c}{\kappa}\left[\overline{U}_2(n;q,u,\varepsilon) + \overline{U}_3(n;q,u,\varepsilon) + \overline{U}_4(n;q,u)\right],$$

where $\overline{U}_2(n;q,u,\varepsilon) = \mathscr{O}(\log(n)n^{-1/2})$, $\overline{U}_3(n;q,u,\varepsilon) = \mathscr{O}(n^{-1/2})$, and $\overline{U}_4(n;q,u) = \mathscr{O}(n^{-1/2})$, with the explicit forms given in the proof.

Note that the values for λ in parts (a) and (b) are chosen to produce the nearly optimal rate of convergence in part (b). In practice there may be better choices and we will typically choose λ via cross-validation. As noted in the discussion following Theorem 1.4.1, we may choose λ, u, or pairs (λ, u) from a finite set of values $\mathscr{W}$. In this case the theorem above can be adjusted to hold simultaneously for all elements of $\mathscr{W}$ by replacing the terms $\log(\varepsilon/3)$ on the right-hand side of the inequality in (a) by $\log(\varepsilon/3) + \log|\mathscr{W}|$ and the terms on the right-hand side of (b) that involve $\log(\varepsilon/4)$, which appear in the $\overline{U}_j$ terms defined in the proof, are replaced by $\log(\varepsilon/4)+\log|\mathscr{W}|$. For example, for fixed $\varepsilon \in (0,1]$ and $u \geq 0$, this adds a term that is $\mathscr{O}(\log|\mathscr{W}|n^{-1/2})$ to the right hand side of (b).

In Theorem 1.4.3, $f_{G,\hat{\rho}_{\lambda,\hat{u}(B,\lambda)}}$ and $f_{G,\hat{\rho}_{\lambda,u}}$ are compared to the best (non-stochastic) models in $\mathscr{F}_\Theta$ with an expected cost no greater than B or $B(\hat{\rho}_{\lambda,u})$, respectively. Additionally, the absence of KL terms in the inequalities allows for a more salient comparison to relevant alternatives. In part (b), for any $u \geq 0$ and $\varepsilon \in (0,1]$, the terms beside $R(\overline{\theta}_u)$ on the right hand-side are collectively $\mathscr{O}(\log(n)n^{-1/2})$. With high probability, the regret of $f_{G,\hat{\rho}_{\lambda,u}}$ gets close to the regret an oracle would obtain choosing the best rule from the subset of $\mathscr{F}_\Theta$ with a target population budget no greater than that of $f_{G,\hat{\rho}_{\lambda,u}}$. The rate $\log(n)n^{-1/2}$ is nearly optimal. For example, in the unconstrained case with $B=\infty$, which corresponds to $u=0$ or $\hat{u}=u^*=0$, Kitagawa and Tetenov (2018) show that $n^{-1/2}$ is the optimal rate for bounds on the expected regret of the empirical welfare maximizer over $\mathscr{F}_\Theta$, provided $\mathscr{F}_\Theta$ has a finite VC-dimension (see the discussion there for more details).

Part (a) has the complication of involving $\hat{u}=\hat{u}(B,\lambda)$ and $u^*=u^*(B,\lambda/2)$ as λ grows with n. The effect of $\hat{u}$ is related to the marginal decrease in

$$R_n(f_{G,\hat{\rho}_{\lambda,\hat{u}(B,\lambda)}})+\lambda^{-1}D_{\mathrm{KL}}(\hat{\rho}_{\lambda,\hat{u}(B,\lambda)},\pi)$$

associated with marginal increases in B as the penalty diminishes (λ increases). The behavior of u^* is related to the marginal decrease in the penalized regret of $f_{G,\rho^*_{\lambda,u^*(B,\lambda/2)}}$ associated with marginal increases in B. Suppose we are unlikely to have large marginal gains in empirical or theoretical penalized regret associated with a marginal increase in B at all or small penalty levels (i.e. as $\lambda \to \infty$). Then (a) implies that, with high probability and for large enough sample sizes, the regret of $f_{G,\hat{\rho}_{\lambda,\hat{u}}}$ is close to the regret that would be obtained by an oracle choosing the best policy from the subset of $\mathscr{F}_\Theta$ with an expected cost in the target population that is less than or equal to B. For example, if $u^*=\mathscr{O}(1)$ and $\hat{u}=\mathscr{O}_p(1)$ as n and λ increase, then the terms on the right-hand side of the inequality in (a) other than $R(\overline{\theta})$ are $\mathscr{O}_p(\log(n)n^{-1/2})$.

We conclude this subsection with remarks regarding implications for the proposed treatment assignment rules. One drawback of starting from a fixed B and utilizing $\hat{u} = \hat{u}(B, \lambda)$ is the absence of a counterpart to Theorem 1.4.1 (b) for the cost $K(f_{G,\hat{\rho}_{\lambda,\hat{u}}})$ when $\hat{u}$ is random. Even when $\hat{u}$ and u^* are well behaved so that Theorem 1.4.3 (a) implies it is likely that $f_{G,\hat{\rho}_{\lambda,\hat{u}}}$ will have regret comparable to the best rules in $\mathscr{F}_\Theta$ with expected cost less than B, this may be achieved with an expected cost greater than B. On the other hand Theorem 1.4.1 (a) with $\rho = \hat{\rho}_{\lambda,\hat{u}}$ yields that with probability at least $1 - \varepsilon$,

$$K\left(f_{G,\hat{\rho}_{\lambda,\hat{u}}}\right) \leq B + \frac{1}{\lambda} D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,\hat{u}}, \pi\right) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_c^2}{8n\kappa^2} + \log\frac{1}{\varepsilon}\right],$$

where we have used the fact that $K_n(f_{G,\hat{\rho}_{\lambda,\hat{u}}}) \leq B$ a.s. under the assumptions of the theorem. When $\lambda = \mathscr{O}(n^{1/2})$, for example, whether or not we have an upper bound that approaches B depends on the behavior of this KL term. Unfortunately, $\hat{u}$ and the KL term above are difficult to analyze in this scenario as $\hat{u}$ is essentially defined implicitly to ensure $K_n(f_{G,\hat{\rho}_{\lambda,\hat{u}}}) \leq B$. It is possible to cross-validate B, for example examining values less than B to try and ensure the expected budget is not violated. The comments regarding extending the high probability bounds to apply simultaneously for multiple values of u can be applied to choices for B as well.

On the whole, the procedure starting with a set of values for u may be more compelling. By Theorem 1.4.1 (b) and the surrounding discussion, for values u in a reasonably sized set $\mathscr{W}$, the values of $K_n(f_{G,\hat{\rho}_{\lambda,u}})$ provide reasonable estimates of $K(f_{G,\hat{\rho}_{\lambda,u}})$, the expected costs of these policies conditional on the rules estimated from the sample. These can be utilized to select u. Alternatively, u can be chosen from $\mathscr{W}$ via cross-validation or by some other method. For example, in the case of pure quantity constraints, it may be possible use data from the target population to select u to achieve the correct (or nearly correct) proportion of treatments assigned in the target population. Theorem 1.4.3 (b) and its extension to hold for all $u \in \mathscr{W}$ simultaneously, then indicate it is likely $R(f_{G,\hat{\rho}_{\lambda,u}})$ for the selected u will be comparable to the best treatment rules in $\mathscr{F}_\Theta$ among those whose target population cost does not exceed that of $f_{G,\hat{\rho}_{\lambda,u}}$. Hence by starting from a set of u values, the policy maker can trace out reasonable estimates of the target population budget horizon. At the same time, the policy selected according to these budget estimates is likely to be the best bang for the buck in that the associated regret gets close to that which an oracle would choose for the same target population cost.

1.4.3 The Majority Vote Treatment Rule

Let $\rho \in \mathscr{P}_\pi(\Theta)$. As mentioned in Section 1.3.2, the non-stochastic majority vote treatment rule $f_{\mathrm{mv},\rho}$ in (1.13) is a close relative of the Gibbs rule $f_{G,\rho}$ that can prove numerically more stable in practice. In the classification literature, it is well known that the risk associated with the majority vote rule, where risk is defined for a zero-one loss function, is upper bounded by twice the risk associated with the Gibbs classification method (e.g., Langford and Shawe-Taylor (2003), McAllester (2003a)). Hence analysis of the Gibbs treatment rule is often used to justify use of the majority vote. Additionally, the "2×" upper bound can be loose and it is not uncommon for majority vote rules to outperform Gibbs rules. We refer to Germain et al. (2015) for further discussion regarding the majority vote versus the Gibbs method for classification settings. Here, we show that, as in the classification setting, the majority vote treatment rule can inherit desirable qualities from the Gibbs treatment rule in the budget constrained treatment rule setting.

While the majority vote rule $f_{\mathrm{mv},\rho}$ is not guaranteed to satisfy the same budget as its Gibbs counterpart $f_{G,\rho}(x)$, we can still show that when $f_{G,\rho}(x)$ is close to $f^*_{B(\rho)}(x)$, the optimal rule for its budget,

$$B(\rho) = K(f_{G,\rho}),$$

then $f_{\mathrm{mv},\rho}$ will also be close to $f^*_{B(\rho)}$. The measurement of closeness, defined shortly, depends on both the welfare achieved and deviations from the budget $B(\rho)$. We will suppose that

$$B(\rho) > E_Q[\delta_c(X)1\{\delta_c(X) < 0\}]. \tag{1.24}$$

That is, $f_{G,\rho}$ does not achieve the exact cost of the cost-minimizing rule $1\{\delta_c(x) < 0\}$ for $x \in \mathscr{X}$.

If (1.24) were an equality, the budget of $f_{G,\rho}$ would be such that a policy maker faced with this budget would need to ignore welfare and seek the lowest cost rule. Hence, when we are interested in maximizing welfare with a budget constraint, it is reasonable to rule out the case where the solution to the policy maker's problem is to ignore welfare and seek the lowest cost. In addition to (1.24), we will assume that $\delta_y(X)$ and $\delta_c(X)$ have bounded densities so that optimal solution to the decision makers in Theorem 1.3.1 is deterministic.

Under (1.24), Assumption 1.3.1, and the condition that $\delta_y(X)$ and $\delta_c(X)$ have bounded densities, Theorem 1.3.1 yields that the optimal budget-constrained policy for the budget $B(\rho)$ of the Gibbs rule $f_{G,\rho}$ is of the form

$$f^*_{B(\rho)}(x) = 1\{\delta_y(x) - \eta_{B(\rho)}\delta_c(x) > 0\},\ \ x \in \mathscr{X}, \tag{1.25}$$

for a constant $\eta_{B(\rho)}$. It also follows from Theorem 1.3.1 that either $\eta_{B(\rho)} = 0$ and $K(f^*_{B(\rho)}) < B(\rho)$ or else $\eta_{B(\rho)} > 0$ and $K(f^*_{B(\rho)}) = B(\rho)$. Recalling the definition of the welfare-regret under a budget constraint in (1.11),

$$R_{B(\rho)}(f) \equiv W\left(f^*_{B(\rho)}\right) - W\left(f\right),$$

it is clear that $R_{B(\rho)}(f_{G,\rho})$ is non-negative. It is small only when $f_{G,\rho}$ attains a welfare that is close to the budget optimal rule in its own budget class. We will show that when $R_{B(\rho)}(f_{G,\rho})$ is small, $f_{\mathrm{mv},\rho}$ has similar welfare to the optimal policy $f^*_{B(\rho)}$ and is unlikely to violate the budget $B(\rho)$ by a large amount.

First note that if a decision maker faced a budget of $B(\rho)$, it would be reasonable to seek a rule $f : \mathscr{X} \to [0,1]$ that minimizes

$$L_{B(\rho)}(f) \equiv E_Q\left[\left(\delta_y(X) - \eta_{B(\rho)}\delta_c(X)\right)\left(f^*_{B(\rho)}(X) - f(X)\right)\right],$$

with the associated loss function

$$\begin{aligned}\ell_{B(\rho)}(f,x) &= \left(\delta_y(x) - \eta_{B(\rho)}\delta_c(x)\right)\left(f^*_{B(\rho)}(x) - f(x)\right) \\ &= \begin{cases} 0 & \text{if } f^*_{B(\rho)}(x) = f(x), \\ \left|\delta_y(x) - \eta_{B(\rho)}\delta_c(x)\right| & \text{if } f^*_{B(\rho)}(x) \neq f(x). \end{cases}\end{aligned}$$

By the form of $f^*_{B(\rho)}$ in (1.25), $L_{B(\rho)}(f)$ is non-negative and attains the value zero only when $f(X) = f^*_{B(\rho)}(X)$ almost surely. Of course, such a loss function cannot yield an estimation strategy directly because δ_y, δ_x, and $\eta_{B(\rho)}$ are unknown. However, when $L_{B(\rho)}(f)$ is small, this means we are unlikely to encounter a set of co-variates X for which f assigns treatment and $\eta_{B(\rho)}\delta_c(X)$ exceeds $\delta_y(X)$ by a large amount. We have the following result

Theorem 1.4.4 *Let $\rho \in \mathscr{P}_\pi(\Theta)$. Let Assumptions 1.3.1 and 1.3.2 hold and also assume that* (1.24) *holds and $\delta_c(X)$ and $\delta_y(X)$ have bounded densities so that $E_Q[1\{\delta_y(X) = \eta_{B(\rho)}\delta_c(X)\}] = 0$. Then*

$$L_{B(\rho)}\left(f_{\text{mv},\rho}\right) \leq 2R_{B(\rho)}\left(f_{G,\rho}\right).$$

We note that the expectation in the definition of $L_{B(\rho)}(f)$ is taken with respect to a draw from the target population. When ρ is dependent on the sample data, the result and proof still hold, conditional on the estimated rule or sample, provided that (1.24) can be assumed to hold almost surely for ρ or with high probability if considering probabilistic bounds such as those in Sections 1.4.1 and 1.4.2. This is reasonable to assume for $\hat{\rho}_{\lambda,u}$, particularly when u is not so large that no treatments will be assigned. The notion that, for appropriately chosen values of λ, $R_{B(\hat{\rho}_{\lambda,u})}(f_{G,\hat{\rho}_{\lambda,u}})$ is small is exactly the implication of Theorems 1.4.2 (b) and 1.4.3 (b).

For example, assume that the conditions of Theorem 1.4.3 hold, take $\lambda = \kappa\sqrt{nq}/(M_y + uM_c)$ (although we continue to write λ to reduce clutter in the notation), and suppose that (1.24) holds almost surely for $\rho = \hat{\rho}_{\lambda,u}$ and that $\delta_c(X)$ and $\delta_y(X)$ have bounded densities. Then by

Theorem 1.4.3 (b), with probability at least $1-\varepsilon$ it holds that

$$\begin{aligned}
&R_{B(f_{G,\hat{\rho}_{\lambda,u}})}\left(f_{G,\hat{\rho}_{\lambda,u}}\right) \\
&\leq \operatorname*{arg\,min}_{\theta \in \Theta_{B(\hat{\rho}_{\lambda,u})}} \left[R_{B(f_{G,\hat{\rho}_{\lambda,u}})}(\theta)\right] + \frac{M_y + uM_c}{\kappa}\left[\overline{U}_2(n;q,u,\varepsilon) + \overline{U}_3(n;q,u,\varepsilon) + \overline{U}_4(n;q,u)\right]
\end{aligned}$$

where we have done some simple algebra on the inequality of part (b) of Theorem 1.4.3 utilizing the definitions of regret and regret under a budget constraint. The above also uses the notation

$$R_{B(f_{G,\hat{\rho}_{\lambda,u}})}(\theta) = W\left(f^*_{B(f_{G,\hat{\rho}_{\lambda,u}})}\right) - W\left(f_\theta\right).$$

Recall that the terms outside of the arg min on the right-hand side of the above inequality are at most $\mathscr{O}(\log(n)n^{-1/2})$ for fixed $u \geq 0$, $q \in \mathbb{N}$ and $\varepsilon \in (0,1]$. If, for example,

$$\delta_y(X) = \phi(X)^\intercal \theta_y, \text{ and } \delta_c(X) = \phi(X)^\intercal \theta_c,$$

for some $\theta_y, \theta_c \in \mathbb{R}^q$, then we would have

$$\operatorname*{arg\,min}_{\theta \in \Theta_{B(\hat{\rho}_{\lambda,u})}} \left[R_{B(f_{G,\hat{\rho}_{\lambda,u}})}(\theta)\right] = 0.$$

In this case, the above combined with Theorem 1.4.4 produce that, with probability at least $1-\varepsilon$, $L_{B(\hat{\rho}_{\lambda,u})}(f_{\mathrm{mv},\hat{\rho}_{\lambda,u}})$ is bounded above by terms that are $\mathscr{O}(\log(n)n^{-1/2})$.

1.5 Simulation Study and Implementation Details

In this section we evaluate the proposed treatment assignment methodology in a simulation environment. We also discuss model estimation and implementation. Section 1.5.1 describes the simulation environment and findings. Section 1.5.2 describes a model estimation strategy using the Sequential Monte Carlo (SMC) approach and discusses the implementation choices utilized in the simulation.

1.5.1 Simulation Study

We assign treatments utilizing $\hat{\rho}_{\lambda,u}$ in the following simulation environments. We take $X = (X_1, X_2, X_3)$ where $X_j \sim \text{Unif}(-1,1)$ for $j = 1,2,3$ are i.i.d. uniform random variables. Letting $\Lambda(v) = (1+\exp(-v))^{-1}$ denote the logistic function, potential outcomes are determined via

$$Y_d = \max\{X_1 + X_2, 0\} + \max\{X_3, 0\} + 4d\Lambda\left(2\left(X_1 + X_1X_2 + X_2\right)\right) + \varepsilon, \;\; d \in \{0,1\},$$

where ε is taken to be a standard normal random variable that is truncated to take values in $[-2,2]$ and is independent of all other variables considered. Potential costs are determined via

$$C_0 = 0, \;\; C_1 \sim \text{Binom}\left(6, \frac{4\Lambda\left(a(3X_2 + 1.5X_3)\right)}{6}\right),$$

where a is a constant. Lastly, $e(x) = 1/2$ for all $x \in \mathscr{X}$ so that $D \sim \text{Bern}(1/2)$ and is independent of the other variables. We consider $a \in \{1,2,4\}$.

Each choice of a corresponds to a different data generating process (DGP) and for each we perform the following simulation study separately. We simulate training sets each with sample size $n = 1,000$. A testing sample of size $n_{\text{test}} = 10,000$, which is re-used across training sample iterations, yields approximately the true costs and benefits from of any considered treatment rule. We consider 100 training simulation replicates. Using knowledge of the DGP, we can calculate $E_Q[Y_{1,i} - Y_{0,i}|X_i]$ and $E_Q[C_{1,i}|X_i]$ for each testing set observation. Then, for a rule $f(x) : \mathscr{X} \to [0,1]$, we use the testing set to obtain the (approximate) gain and cost associated with f,

$$\text{Gain of } f = E_Q\left[(Y_1 - Y_0)\, f(X)\right],$$

and

$$\text{Cost of } f = E_Q \left[C_1 f(X) \right].$$

The Gain of f is the expected increase in welfare, relative to no treatments, associated with the treatment rule policy while the Cost of f is its cost.

Section 1.5.2 describes the Sequential Monte Carlo procedure used to sample from $\hat{\rho}_{\lambda,u}$ to implement the associated Gibbs or majority vote rule. We consider values of u increasing from 0 to 2 in increments of 0.05. For each choice of u, λ is chosen by 4-fold cross-validation to maximize $W_n(f) - uK_n(f)$ across hold-out folds, where $f = f_{G,\hat{\rho}_{\lambda,u}}$ for the Gibbs rules and $f = f_{\text{mv},\hat{\rho}_{\lambda,u}}$ for the majority vote rules. We thus obtain treatment rules with varying gain-cost pairs for different choices of u and can obtain cross-validation-based estimates of these pairs during the estimation stage.

To make estimation from a training sample operational, we must specify a treatment rule space $\mathscr{F}_\Theta$ and prior π. With d_x denoting the dimension of $\mathscr{X}$ ($d_x = 3$ in the simulation setting), for $k \in \mathbb{N}$ and $q_k = \binom{d_x+k}{k}$, the polynomial transformation on $\mathscr{X}$ of order at most k is defined as

$$\mathscr{F}_\Theta^{\text{poly}}(k) = \left\{ m(x) : m(x) = \sum_{j=1}^{q_k} \theta_j \phi_j(x), \theta \in \mathbb{R}^{q_k} \right\}, \tag{1.26}$$

where the summation is over all monomials $\phi_j(x) = \prod_{\ell=1}^{d} x_\ell^{p_{j\ell}}$ with $\sum_{\ell=1}^{d} p_{j\ell} \leq q$, $p_{j\ell} \in \mathbb{N} \cup \{0\}$. We take $\mathscr{F}_\Theta$ to be the family of rules described in (1.9) and (1.10) where the transformations $\phi_j(x)$ are the monomials used in the construction of the polynomial transformations on $\mathbb{R}^3$ of order at most 2 with the monomials normalized by their sample mean and standard deviation calculated from training data. We set π to be the standard multivariate normal distribution over $\mathbb{R}^{10}$.

As an alternative treatment rule, we consider the approach of Sun et al. (2021). Under our simulation setting, where for example $C_0 \leq C_1$ almost surely, f_B^* in Theorem 1.3.1 takes the form

$$f_B^*(x) = 1\left\{\frac{\delta_y(x)}{\delta_c(x)} > \eta_B\right\},$$

for some constant η_B. Sun et al. (2021) show that $\delta_y(x)/\delta_c(x)$ can be estimated nonparametrically by re-purposing the so-called generalized random forest methodology of Athey et al. (2019). When costs can be observed at the time of treatment assignment, their approach first estimates $\delta_y(x)/\delta_c(x)$ for $x \in \mathscr{X}$. This produces an estimate of the conditional welfare to conditional cost ratio $\delta_y(X_i)/\delta_c(X_i)$ for each observation in the target group[5]. These ratio estimates are ranked according in descending order and treatments are allotted according to this order until the budget is exhausted. We call such a method of assignment, where a ranking is derived for members of the target group who are then treated in that order until the budget is reached, a “batch implementation” method. Additionally, as a baseline rule, we estimate the CATE $\delta_y(x) = E_Q[Y_1 - Y_0|X = x]$ using the generalized random forest of Athey et al. (2019) and then use the resulting scores in the target group for a batch implementation. This baseline approach does not factor costs into the treatment decisions. In our simulations, these methods are implemented using R 4.2.2 (R Core Team (2023)) with the *grf* package (Tibshirani et al. (2022)) following the described adaptation in Sun et al. (2021) for their approach. The default package settings were except that the known treatment probabilities supplied to the algorithm.

The approach of Sun et al. (2021) and the baseline that ignores cost utilize batch implementations while the Gibbs and majority vote methods do not. To compare like-for-like, the majority vote models associated with $\hat{\rho}_{\lambda,u}$ for a range of u values (with λ chosen via cross-validation for each u) are amenable to a batch implementation method. An algorithm for implementing a batch treatment rule utilizing the majority vote rules is described below.

[5] By target group we mean individuals or units for whom treatment assignment must be determined, typically this is the wider population from which the sample comes from that consists of individuals or units not used in fitting treatment rules.

Batch treatment implementation utilizing majority vote rules

Input Target group observations indexed by $\mathscr{I}_{\text{target}} = \{1, 2, \ldots, n_{\text{target}}\}$ with n_{target} total observations and covariates $\{X_j : j \in \mathscr{I}_{\text{target}}\}$, minimum cost $B_{\min}$, number of bins used denoted n_{bin}, budget B, set of u values denoted $\mathscr{W}_u$, majority vote rules $f_{\text{mv},\hat{\rho}_{\tilde{\lambda}_u,u}}$ for each $u \in \mathscr{W}_u$ along with cost estimates $\widehat{\text{cost}}(f_{\text{mv},\hat{\rho}_{\tilde{\lambda}_u,u}})$. If treatment is assigned to X_j, we then observe the cost of treating individual j, $C_{1,j}$.

Output $\mathscr{I}_{\text{treat}} \subseteq \mathscr{I}_{\text{target}}$, a subset of individuals in the target group assigned treatment.

Step 1: Initialization

Set $B_0 \leftarrow B_{\min}$ and $\mathscr{I}_{\text{treat}} \leftarrow \emptyset$.

Step 2: Treatment determinations

For $i = 1 : n_{\text{bin}}$

- Set $u_i \leftarrow \underset{u \in \mathscr{W}_u}{\arg\min} \left| \widehat{\text{cost}}\left(f_{\text{mv},\hat{\rho}_{\tilde{\lambda}_u,u}}\right) - \left(\frac{i(B - B_{\min})}{n_{\text{bin}}}\right) \right|$.
- Let $\mathscr{I}_i = \{\alpha_i(1), \alpha_i(2), \ldots\}$ denote the ordered ranking of target group observations not currently in the set $\mathscr{I}_{\text{treat}}$ in decreasing order of the majority vote scores. That is, in decreasing order of $\int_{\Theta} f_{\theta}(X_j) d\hat{\rho}_{\tilde{\lambda}_{u_i},u_i}(\theta)$ for $j \in \mathscr{I}_{\text{target}} \cap \mathscr{I}^{\mathbf{c}}_{\text{treat}}$. For example, $\alpha_i(1)$ gives the index of the individual with the largest such majority vote score that is in $\mathscr{I}_{\text{target}}$ but not currently in $\mathscr{I}_{\text{treat}}$, provided that $\mathscr{I}_{\text{target}} \cap \mathscr{I}^{\mathbf{c}}_{\text{treat}} \neq \emptyset$. In the latter case, $\mathscr{I}_i = \emptyset$.
- Set $k \leftarrow 1$.
- **While** $B_0 < B \times n_{\text{target}}$ **and** $k \leq |\mathscr{I}_i|$ **do** $\mathscr{I}_{\text{treat}} \leftarrow \mathscr{I}_{\text{treat}} \cup \alpha_i(k)$, $B_0 \leftarrow B_0 + C_{1,\alpha_i(k)}$, and then $k \leftarrow k + 1$.

End For

The algorithm above divides the cost space below the budget into bins and then performs a batch implementation at each bin using the majority vote scores of the model with an estimated cost nearest to that bin's endpoint. Note that we are using the notation $\tilde{\lambda}_u$ in $f_{\text{mv},\hat{\rho}_{\tilde{\lambda}_u,u}}$ to reflect that λ varies with u and is data dependent. For $\hat{\text{cost}}(f_{\text{mv},\hat{\rho}_{\tilde{\lambda}_u,u}})$, one could use $K_n(f_{\text{mv},\hat{\rho}_{\tilde{\lambda}_u,u}})$, an estimate of the cost arising during the cross-validation of λ, or some other estimate such as one arising from an auxiliary testing dataset if one is available. Minor modifications may improve the performance, for example dropping any values of u from consideration in Step 2 if there exists another u' with a corresponding estimated majority vote model that has lower estimated cost but higher estimated welfare. However, in our simulations we use the simpler version presented above. We take $\hat{\text{cost}}(f_{\text{mv},\hat{\rho}_{\tilde{\lambda}_u,u}})$ to be the average cost associated with $f_{\text{mv},\hat{\rho}_{\tilde{\lambda}_u,u}}$ across the hold-out fold samples during the cross-validation of λ when estimating $\hat{\rho}_{\lambda,u}$ for the majority vote model.

The batch implementation utilizing the majority vote rules is noteworthy because, when batch implementation is feasible, it controls costs accurately. In our simulations, we created 20 equally spaced cost bins, starting at 0 and with endpoints increasing from 0.1 to 2 by increments of 0.1. We treated each end point as a desired budget level and applied the batch implementation that utilizes the majority vote models. For example, the first desired budget level is $B = 0.1$ and utilizes $n_{\text{bin}} = 1$ in the algorithm above, while the last desired budget is $B = 2$ and we set $n_{\text{bin}} = 20$. Throughout, we take $B_{\min} = 0$. For each budget level we also applied the alternative batch implementation methods. The gains associated with models fit to each training sample iteration were calculated using the test set. Then these gains were averaged over all training sample iterations to produce Figures 1.5, 1.7 and 1.9 for $a = 1,2,4$, respectively. We denote the batch implementation method utilizing the majority vote models by "PB-B", we denote the non-parametric method of Sun et al. (2021) centered around the conditional welfare to conditional cost ratio by "R-NP", and we denote the baseline that ignores cost by "Ignore Cost" or IC in subsequent discussion.

We will refer to the non-batch-implemented stochastic Gibbs and non-stochastic majority vote methods by “PB-G” and “PB-MV”, respectively. To assess these methods, we utilize “cost curves” to compare the gain-cost trade-off of the considered rules at different budget levels. These are constructed as follows. For a single training sample iteration, for each u we estimate a Gibbs rule and a majority vote rule. We then evaluate the true cost and gain associated with these treatment rules (for different choices of u) using the test data. Once we have the true costs associated with these rules, we estimate the R-NP ratios and IC CATE scores from the training sample and implement these rules via batch implementation in the testing data. For each u choice and for each PB-MV and PB-G rule, the R-NP and IC rules are implemented to stop assigning treatment when they reach the same cost as the PB-MV or PB-G rule of interest. In this way we are comparing models with the same true costs.

For each training sample, the various (approximately) true gain-cost points associated with different u choices for the PB-MV and PB-G methods are plotted in gain-cost space along with the associated points for the R-NP and IC models. The gain-cost curve for the iteration is then estimated by interpolating between these points. For a single training sample iteration, this process is illustrated in Figures 1.11 and 1.12 for the DGP with $a = 1$. Then, the gain-cost curves for all training sample iterations are averaged (vertically) to produce the final (approximately) true gain-cost curves. This procedure for the DGP with $a = 1$ then produces Figure 1.6. The black lines in these figures give the gain-cost pairs that would result from randomly assigning treatment in the target population until the particular cost level is achieved. The cost curves for the DGPs with $a = 2$ and $a = 4$ are presented in Figures 1.8 and 1.10, respectively.

We can now discuss the main takeaways and results from the simulation study. Figures 1.5-1.12 present the cost curves from the simulation study while Table 1.1 collects select data points from these graphs for a more precise snapshot. For $a = 1$, the PAC-Bayesian methods PB-G, PB-MV, and PB-B perform quite closely to the R-NP method. In this setting, the R-NP slightly outperforms the PB-G and PB-MV methods across most cost levels, with the gap in out-performance slightly increasing at greater cost levels. The PB-B models, on the other hand, perform quite similarly to the R-NP method across the cost levels in this setting, with ± 0.01 differences in welfare at a few cost levels.

As a increases to 2 and 4, all of the PAC-Bayesian-based rules improve their performance relative to the R-NP approach. PB-B rules yield higher welfare gains than the R-NP rules at lower to middle cost levels while slightly lagging the welfare of the R-NP models at higher cost levels for $a = 2$ and slightly out-performing them at $a = 4$. The out-performance of PB-B models increases slightly at lower cost levels as a increases from 2 from 4. For $a \in \{2,4\}$, relative to the R-NP models, the PB-MV and PB-G models now yield higher welfare gains at lower cost levels, perform similarly at middling cost levels, and are slightly beaten at the highest budgets. As the cost/budget level increases, the optimal rules in these simulation environments involve treating a higher proportion of the target population, eventually treating everyone as cost levels are allowed to rise enough. The optimization problem that the Gibbs posterior solves is penalized towards allowing a degree of randomness in the resulting Gibbs rule (see, for example, the $D_{\mathrm{KL}}(\rho, \pi)$ term in (1.15)). This could help to explain why the PB-G and closely related PB-MV models lag slightly at the highest cost levels whereas the PB-B implementation that treats until the budget is met performs well at these levels.

Note that

$$\delta_y(x) = 4\Lambda\left(2\left(x_1 + x_1 x_2 + x_2\right)\right), \ \delta_c(x) = 4\Lambda\left(a(3x_2 + 1.5x_3)\right).$$

For values of v near zero, $\Lambda(v)$ is approximately linear in v and so the above compositions are also approximately linear in $x_1, x_1 x_2$, x_2 and x_3 near the origin. For values further from the origin, which are encountered with increasing probability as a increases, this linear approximation worsens. When we are likely to observe combinations of X_1, X_2, X_1X_2 and X_3 that are further from the origin, the conditional welfare to cost ratio in the optimal rules is a more complex object in these regions that is less well approximated by individual rules in $\mathscr{F}_\Theta$ and has increasing variance as a increases. This simulation study could suggest the PAC-Bayesian approaches may have benefits over the R-NP method when conditional expected costs are noisier.

In practice it is desirable to compare alternative methods prior to implementation. For example, via evaluation using an auxialry testing data set separate from that which the models are trained on. One drawback of the approach of Sun et al. (2021) and other batch implementation methods is they cannot be evaluated in a traditional way using test data withheld from model estimation. For example, test sample data points that a batch implementation method may rank highly for treatment may not have received the treatment and thus we do not observe the costs accruing properly to know when a batch implementation method would stop assigning treatments. We note that Sun et al. (2021) is a working paper and since this paper was started the authors have added material aimed at addressing this issue.

It also is important to note that there are a number of settings where the forest-based R-NP method is not viable whereas the PAC-Bayesian approaches considered here remain applicable. Batch implementations are not always viable. The cost of a treatment may not be realized until sometime after treatment assignment and one may not always have the full target group available when the rule must be set. Batch implementations, where treatment is assigned until the budget is hit, could also be unacceptable to policy makers in settings where the "budget" is something with a negative connotation like a complication rate in a medical setting.

Additionally, the R-NP rule can only be applied when $C_0 \leq C_1$ a.s., which rules out certain circumstances relevant to policy makers. For example, as noted in Sun (2021), Hendren and Sprung-Keyser (2020) identify fourteen welfare programs out of 133 considered that are estimated to have negative or zero net cost to the government. The EWM based approach of Sun (2021) can accommodate the setting where $C_0 > C_1$ with positive probability, as can the PB-G and PB-MV methods considered here. However, the approach of Sun (2021) may be difficult to implement when allowing for more flexible decision rule classes (she considers threshold rules that vary with a covariate in her application) and lacks the budget efficiency properties derived here. An additional benefit of the PAC-Bayesian approaches here is their ability to utilize estimation tools from the Bayesian literature as demonstrated in Section 1.5.2 below. Lastly, while one could estimate $\delta_y(x)$ and $\delta_c(x)$ separately and try to build a workaround via Theorem 1.3.1 when $C_0 > C_1$ is possible, the resulting ratio estimates may have increased variance and will again require batch implementation, which adds a complication in this setting.

1.5.2 Implementation and Estimation via Sequential Monte Carlo

To implement treatment rules associated with $\hat{\rho}_{\lambda,u}(\theta)$, we must evaluate the treatment assignment probabilities or majority vote scores of the form

$$\int_{\Theta} f_\theta(x) d\hat{\rho}_{\lambda,u}(\theta), \;\; x \in \mathscr{X}. \tag{1.27}$$

To do so, we utilize the Sequential Monte Carlo (SMC) procedure considered, for example, in Del Moral et al. (2006). While a Markov Chain Monte Carlo (MCMC) approach also could be derived, recently Ridgway et al. (2014) and Alquier et al. (2016) have highlighted the usefulness of the SMC procedure in PAC-Bayesian applications. One benefit is the ability to sample from a sequence of Gibbs posterior distributions for a range of λ values. This can ease the computational burden for cross-validation. Here we discuss key elements of the approach, provide an estimation algorithm for our setting, and discuss implementation. We also discuss the choices utilized in implementing the procedure for Section 1.5.1.

Throughout, we make the following computational adjustment to the definition of $\hat{\rho}_{\lambda,u}$ in order to make the implementation choices for Section 1.5.1 applicable to more general settings. We define $\hat{\rho}_{\lambda,u}$ to be the distribution over Θ with RN derivative with respect to π given by

$$\frac{d\hat{\rho}_{\lambda,u}}{d\pi}(\theta) = \frac{\exp\left[-\lambda\left(u\overline{K}_n(\theta) - \overline{W}_n(\theta)\right)\right]}{Z(\lambda,u)}, \tag{1.28}$$

where

$$Z(\lambda,u) = \int_{\Theta} \exp\left[-\lambda\left(u\overline{K}_n(\theta) - \overline{W}_n(\theta)\right)\right] d\pi(\theta),$$

and

$$\overline{W}_n(\theta) = \frac{W_n(\theta)}{\frac{1}{n}\sum_{i=1}^n \delta_{y,i}} \text{ and } \overline{K}_n(\theta) = \frac{K_n(\theta)}{\frac{1}{n}\sum_{i=1}^n \delta_{y,i}}.$$

This adjustment is relevant when the average treatment effect is expected to be is positive. Clearly, if we denote $\hat{\delta}_y = n^{-1}\sum_{i=1}^n \delta_{y,i}$, the distribution $\hat{\rho}_{\lambda,u}$ in (1.28) is equivalent to $\hat{\rho}_{(\hat{\delta}_y\lambda),u}$ in Definition 1.3.3. In practice we choose λ via cross-validation from a wide range of values.

For given choices of $\lambda > 0$ and $u \geq 0$, the SMC algorithm we adopt samples from $\hat{\rho}_{\lambda,u}$ to evaluate (1.27) by simulating a set of parameter draws from each of a sequence of distributions $\{\hat{\rho}_{\lambda_t,u}\}_{t=0}^T$. Here,

$$0 = \lambda_0 < \lambda_1 < \cdots < \lambda_T = \lambda$$

is an increasing temperature ladder that must be specified. Note that $\hat{\rho}_{\lambda_0,u} = \pi$, which the user may specify and we assume can be sampled from. The temperature ladder $\{\lambda_t\}_{t=0}^T$ is intended to be such that the corresponding distributions $\hat{\rho}_{\lambda_t,u}$ progress gradually from π to the target distribution $\hat{\rho}_{\lambda,u}$.

For each $t = 0,\ldots,T$, the SMC algorithm produces a set of N weighted samples $\{\Psi_t^{(i)}, \theta_t^{(i)}\}_{i=1}^N$ with $\Psi_t^{(i)} > 0$ and $\sum_{i=1}^N \Psi_t^{(i)} = 1$ where $\theta_t^{(i)} \in \Theta$ for all t and i in our setting. The set of parameter draws $\{\theta_t^{(i)}\}_{i=1}^N$ are referred to as particles (there are N weighted particles for each t). SMC combines MCMC moves with sequential importance sampling; we refer to Del Moral et al. (2006) for additional details and discussion. This produces weighted particles that emulate, in terms of computing expectations, samples from the distributions $\hat{\rho}_{\lambda_t,u}$ associated with

$$\frac{d\hat{\rho}_{\lambda_t,u}}{d\pi}(\theta) = \frac{\exp\left[-\lambda_t\left(u\overline{K}_n(\theta) - \overline{W}_n(\theta)\right)\right]}{Z_t},\ \ Z_t = \int_\Theta \exp\left[-\lambda_t\left(u\overline{K}_n(\theta) - \overline{W}_n(\theta)\right)\right] d\pi(\theta).$$

Conditional on $\hat{\rho}_{\lambda_T,u}$, under general conditions, for a $\hat{\rho}_{\lambda_T,u}$-integrable function $\varphi : \Theta \to \mathbb{R}$,

$$\sum_{i=1}^N \Psi_T^{(i)} \varphi\left(\theta_T^{(i)}\right) \overset{a.s.}{\to} \int_\Theta \varphi(\theta)\, d\hat{\rho}_{\lambda_T,u}(\theta) \text{ as } N \to \infty.$$

In our setting, we are interested in $\varphi(\theta) = f_\theta(x)$ to approximate (1.27) via

$$\sum_{i=1}^{N} \Psi_T^{(i)} f_{\theta_T^{(i)}}(x), \;\; x \in \mathscr{X}.$$

Once we have run the SMC algorithm to yield $\{\Psi_T^{(i)}, \theta_T^{(i)}\}_{i=1}^N$ for a given pair $(\lambda, u) = (\lambda_T, u)$, the treatment probability or majority vote score for any value x in the covariate space can be computed as above. Alternatively, for example, we may be interested in $\varphi(\theta) = K_n(\theta)$, to approximate $K_n(f_{G,\hat{\rho}_{\lambda,u}})$.

The SMC algorithm utilized to estimate the treatment rules in Section 1.5.1 is detailed in the algorithm tables below. We set the input parameters τ_{ESS} and N there equal to $1/2$ and $1,000$, respectively. τ_{ESS} is an Effective Sample Size threshold criterion. When the variance of the weights at a given step t is too high, the SMC procedure utilizes a re-sampling step. This is referred to in Step 2 of the algorithm below. In our application we utilize systematic resampling, which is also outlined below. The choice of temperature ladder, additional algorithm details, and cross-validation points are detailed below the algorithm descriptions.

Tempering SMC Algorithm

Input N (number of particles), $\tau_{\text{ESS}} \in (0,1)$ (ESS threshold), $\{\lambda_t\}_{t=1}^T$ (temperature ladder).

Output $\{\Psi_t^{(i)}, \theta_t^{(i)}\}_{i=1}^N$ for $t = 0, \ldots, T$.

Step 1: initialization

- Set $t \leftarrow 0$. For $i = 1, \ldots, N$, draw $\theta_0^{(i)} \sim \pi$ and set $\Psi_0^{(i)} \leftarrow 1/N$.

Iterate steps 2 and 3

Step 2: Resampling

- If
$$\left\{\sum_{i=1}^{N}\left(\Psi_t^{(i)}\right)^2\right\}^{-1} < \tau_{\mathrm{ESS}}N,$$
resample $\left\{\Psi_t^{(i)},\theta_t^{(i)}\right\}_{i=1}^{N}$ yielding equally weighted resampled particles $\left\{\frac{1}{N},\overline{\theta}_t^{(i)}\right\}_{i=1}^{N}$ and set $\left\{\Psi_t^{(i)},\theta_t^{(i)}\right\}_{i=1}^{N} \leftarrow \left\{\frac{1}{N},\overline{\theta}_t^{(i)}\right\}_{i=1}^{N}$. Otherwise, leave $\left\{\Psi_t^{(i)},\theta_t^{(i)}\right\}_{i=1}^{N}$ unaltered.

Step 3: Sampling

- Set $t \leftarrow t+1$; if $t = T+1$, stop.

- For $i = 1,\ldots,N$, draw $\theta_t^{(i)} \sim K_t(\theta_{t-1}^{(i)},\cdot)$, where K_t is an MCMC kernel with invariant distribution $\hat{\rho}_{\lambda_t,u}$, and evaluate the unnormalized importance weights
$$\omega_t^{(i)}\left(\theta_{t-1}^{(i)}\right) = \exp\left[\lambda_{t-1}\left(u\overline{K}_n\left(\theta_{t-1}^{(i)}\right) - \overline{W}_n\left(\theta_{t-1}^{(i)}\right)\right) - \lambda_t\left(u\overline{K}_n\left(\theta_{t-1}^{(i)}\right) - \overline{W}_n\left(\theta_{t-1}^{(i)}\right)\right)\right].$$

- For $i = 1,\ldots,N$, set
$$\Psi_t^{(i)} \leftarrow \frac{\Psi_{t-1}^{(i)}\omega_t\left(\theta_{t-1}^{(i)}\right)}{\sum_{j=1}^{N}\Psi_{t-1}^{(j)}\omega_t\left(\theta_{t-1}^{(j)}\right)}.$$

Resampling Algorithm (systematic resampling):

Input A set of (normalized) weights and associated particles, $\left\{\Psi_t^{(i)},\theta_t^{(i)}\right\}_{i=1}^{N}$ for some $t \in \{0,\ldots,T\}$.

Output Resampled particles for equal weighting, $\left\{\overline{\theta}_t^{(i)}\right\}_{i=1}^{N}$

- Draw $u \sim U\left[0,\frac{1}{N}\right]$.

- Compute cumulative weights $C^{(i)} = \sum_{m=1}^{i}\Psi_t^{(m)}$ for $i = 1,...,N$.

- Set $m \leftarrow 1$.

```
• For $i = 1 : N$
      While $u < C^{(i)}$ do $\overline{\theta}_t^{(m)} \leftarrow \theta_t^{(i)}$.
      $m \leftarrow m + 1$, and $u \leftarrow u + 1/N$.
  End For
```

Some additional implementation details utilized in Section 1.5.1 are as follows. For the MCMC kernel in the sampling step of the SMC algorithm, we use a Gaussian random-walk Metropolis kernel with covariance matrix proportional to the empirical covariance matrix of the current set of particles. We scale the empirical covariance of the step t particles by $t^{-0.9}$. In practice, the scaling factor can be adjusted, possibly dynamically rather than with a general rule like $t^{-0.9}$, to ensure reasonable acceptance rates in the MCMC steps of the SMC algorithm.

For a temperature ladder, we set $T = 800$ and utilize a piece-wise linear structure as follows. $\{\lambda_t\}_{t=0}^{200}$ increases from 0 to $4/(|1+u|)$ in equally spaced steps. Then, $\{\lambda_t\}_{t=201}^{320}$ increases from $4/(|1+u|)$ to $32/(|1+u|) = 2^5/(|1+u|)$ in equally spaced increments, $\{\lambda_t\}_{t=321}^{470}$ increases from $2^5/(|1+u|)$ to $2^8/(|1+u|)$ in equally spaced increments, and $\{\lambda_t\}_{t=471}^{800}$ increases from $2^8/(|1+u|)$ to $2^{10}/(|1+u|)$ in equally spaced increments. For λ values of interest less than 2^{10}, the temperature ladder is cut short (at fewer than 800 steps) to end once λ_t reaches the desired value. Let $\mathscr{W}_\lambda$ denote the elements of $\{\lambda_t\}_{t=0}^{800}$ that are closest to elements in $\{2^2, (2^2+2^3)/2, 2^3, (2^3+2^4)/2, 2^4, \ldots, 2^{10}\}/(|1+u|)$. Rather than considering each value in $\{\lambda_t\}_{t=0}^{800}$ as a potential choice for a rule, during cross-validation λ chosen from values in $\mathscr{W}_\lambda$.

1.6 Empirical Illustration

Here, we illustrate the procedure for Gibbs treatment rules centered around $\hat{\rho}_{\lambda,u}$ using data from the National Job Training Partnership Act (JTPA) Study. This study has been a popular choice for illustrating individualized treatment rule estimators and is utilized, for example, in

Kitagawa and Tetenov (2018), Mbakop and Tabord-Meehan (2021), and Kitagawa et al. (2023). Detailed descriptions of the study can be found in Orr et al. (1994) and Bloom et al. (1997).

The JTPA study was a randomized controlled trial aimed at assessing the costs and benefits of the training and employment assistance programs of the JTPA. The study randomly assigned each participant to one of two groups. In the first group (the treatment group), participants had access to JTPA services, whereas in the second group (the control group), access to JTPA services was restricted. For example, access was limited to certain services, and a period of time was imposed when a control individual would be ineligible for services. Note that the treatment was ease of access to services, rather than participation in JTPA programs or any other type of compliance. The study collected background information on participants and tracked their earnings in the 30-month period following treatment assignments.

As in Kitagawa and Tetenov (2018), we use an individual's total earnings in the 30 months following treatment as the outcome variable of interest (Y). Our Gibbs treatment rules, like those proposed in the referenced above, are based on two variables that policymakers might consider in designing access policies: an individual's years of education and their earnings in the year prior to treatment assignment. JTPA personnel assessed all participants prior to treatment assignments and provided service type recommendations, which were categorized by Orr et al. (1994) into three types: classroom training, on-the-job training/job search assistance, and other services. To construct our cost variable, we use averages related to these categories, as described below.

Although treatment costs varied between individuals in the study, we don't have exact costs per person. Instead, we define the cost variable C as follows: if an individual received 0 hours of JTPA services, we set their cost to 0. Otherwise, we take their cost to be the average cost of services for individuals with the same gender, treatment assignment, and service recommendation category. These averages are reported in Exhibit 5.3 of Orr et al. (1994) and were adjusted for our purposes to reflect the averages among individuals who received services, using the proportion of individuals in each subcategory who received more than 0 hours of services. As noted in Orr et al. (1994), it is relevant that services utilized correlated with service recommendations and individuals in the control group that did access JTPA services at some point in the study tended to incur similar costs as individuals in the treatment group that received services. The probability of using any services, on the other hand, was greatly impacted by treatment assignment and is a key driver in cost differences.

Our sample consists of 7,675 adults (22 years and older) for whom data on years of education, pre-program earnings, and service hours received are available. As in Kitagawa and Tetenov (2018), we only consider individuals from the original program evaluation and studies around it (e.g. Bloom et al. (1997) and other references provided in Kitagawa and Tetenov (2018)). The probability of being assigned treatment is $2/3$ in this sample. To estimate potential models of interest, we utilize the SMC procedure described in Section 1.5.2. We consider $u \in \{0.05, 0.1, \dots, 3\}$ and cross-validate $\tilde{\lambda}_u \in \{2, 2^2, \dots, 2^{10}\}/(1+u)$ for each. During the cross-validation step, for each u we obtain an estimated cost and welfare, namely the averages of $\hat{B}(\hat{\rho}_{\tilde{\lambda}_u,u})$ and $W_n(f_{G,\hat{\rho}_{\tilde{\lambda},u}})$ across hold out folds. Note we are abusing notation here because, during cross-validation, objects involving $\hat{\rho}_{\tilde{\lambda},u}$ are calculated from the k-fold training sample rather than the entire sample while the cost and welfare estimates are averages of objects calculated using hold out samples.

We use 2-fold cross-validation because the 30-month post-treatment earnings data is highly variable, and there is a potential to overfit noise in smaller cross-validation samples. We take $\mathscr{F}_\Theta$ to be the family of rules described in (1.9) and (1.10), where the transformations $\phi_j(x)$ are the monomials used in the construction of polynomial transformations on $\mathbb{R}^2$ of order at most 3. As in the simulation section, we normalize the monomial transformations by subtracting their sample means and scaling by the sample standard deviations since there is a considerable degree of variation in scale among the transformations, with education taking values from 7 to 18 years and pre-program earnings ranging from \$0 to \$63,000.

The cross-validation-based estimates of the welfare and cost pairs for different values of u are plotted below in Figure 1.2. We dropped points corresponding to models with dominated welfare-cost pairs, i.e., points where there existed an alternative model for a different choice of u with a higher estimated welfare for the same or lower estimated cost, and these are not displayed. Following an approach where we consider multiple values of u and examine feasible budget estimates of the policies, we can see that there are models with average costs per person ranging from roughly \$100 to about \$650. That is, we estimate that when these treatment rules are applied to a wider population similar to the sample, we can achieve average costs per person within these ranges.

We examine three of the estimated models corresponding to the circled points in Figure 1.2 in greater detail. Starting from the left, the first circle corresponds to the model with an estimated cost closest to \$200 and with $u = 2.45$. We refer to this model as the “low-budget model.” The second circled point corresponds to the model with an estimated cost nearest to \$400, with $u = 2.1$, and we call this model the “medium-budget model.” The right circle, which we refer to as the ”no budget model,” has the highest estimated welfare and corresponds to $u = 0.15$. Figure 1.3 presents treatment probabilities for different values of education and pre-program earnings associated with the high, medium, and no budget models. The population densities linked to the covariate space for these models are shown in Figure 1.4.

We observe that treatment probabilities transition fairly smoothly across the covariate space in all three models but less so in the no budget model. Treatment assignment probabilities are not uniform across the covariate space in any of the estimated models. Furthermore, as we consider models with lower costs, we do not simply obtain uniform reductions in treatment probabilities associated with higher-cost models. In the no budget model, treatment probabilities are either 1 or 0 for sizable portions of the covariate space. The no-budget model treats 81% of individuals in the sample on average. The medium-budget model treats 49% of individuals on average, and the treatment probabilities among individuals in the sample range from 46% to 51%. The low-budget model treats 20% of individuals on average, and treatment probabilities among covariate pairs encountered in the sample range from 4% to 66%.

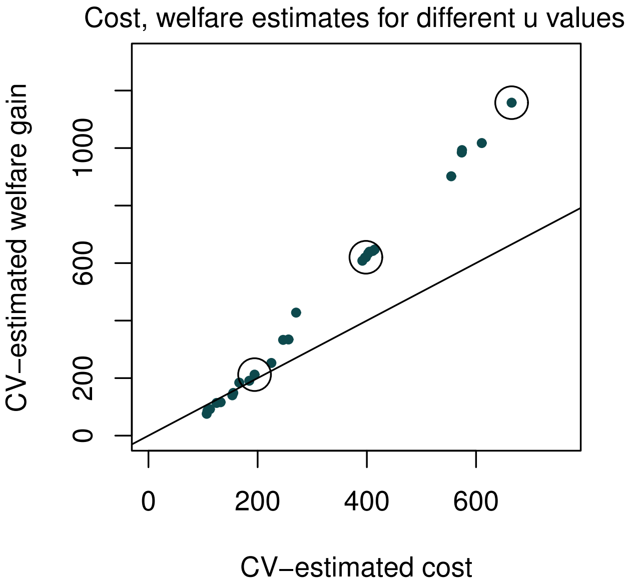


Figure 1.2. Estimated welfare and cost pairs associated with $\hat{\rho}_{\tilde{\lambda}_{u},u}$ for different values of u. The straight line represents points that would have equal cost and welfare. Models with cost estimates closes to \$200 and \$400 are circled, as is the model with the highest estimated welfare gain.

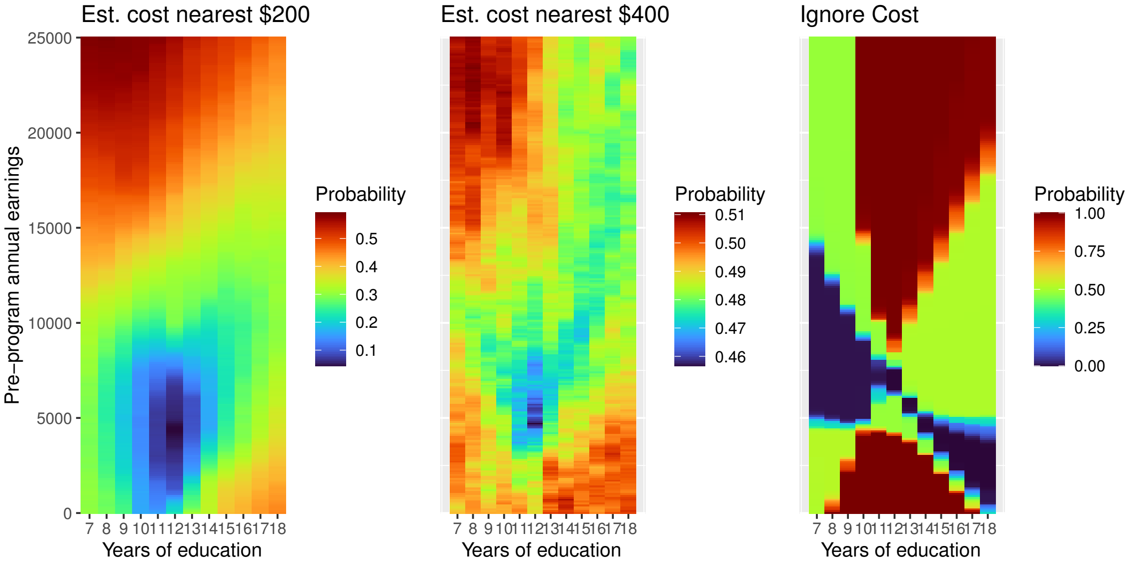


Figure 1.3. Treatment probabilities associated with the low-budget, medium-budget, and no-budget models over different regions of the covariate space.

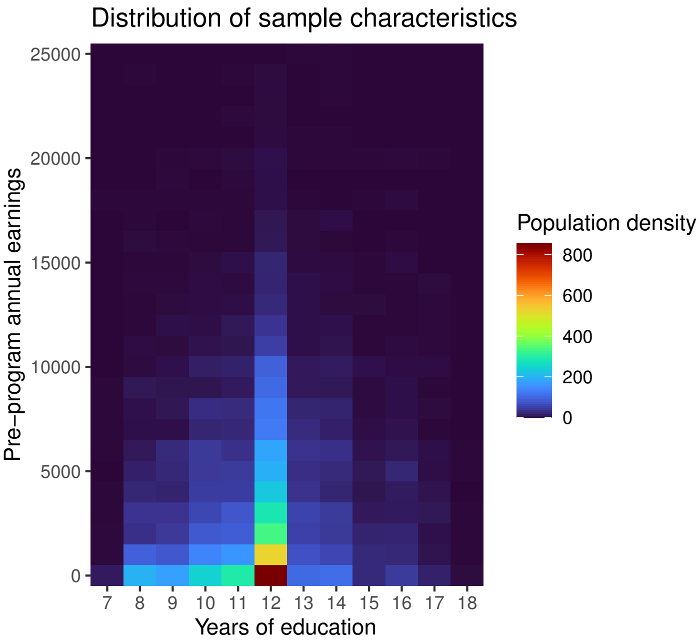


Figure 1.4. Population densities for different regions of the covariate space. The color at each block represents the number of individuals in the sample that fall into the associated education and pre-program earnings level.

One common feature among most of the (non-stochastic) treatment rules estimated in Kitagawa and Tetenov (2018) is the avoidance of treating individuals with the highest levels of education. The most comparable model to those considered here, the no budget model, differs from these rules in that it treats individuals with higher levels of education above a certain income level and randomizes treatment in regions of the covariate space with lower education. It also avoids treating individuals with middling pre-program earnings and low education, although these individuals are relatively less common. As we move to models with lower costs, the medium-budget model becomes more uniform in its treatment probabilities than the other Gibbs rules. It also features increased treatment probabilities among lower and higher education levels compared to the other models, particularly at lower income levels. The low-budget model reduces cost by lowering treatment probabilities among individuals with lower to middle education levels, especially at the most commonly encountered non-zero education levels.

1.7 Conclusion

In this paper, we proposed a new approach to estimating treatment rules in a budget constrained setting. Utilizing the PAC-Bayesian framework, theoretical properties of interest were derived, including generalization bounds and oracle-type inequalities demonstrating a type of budget efficiency for a proposed class of stochastic treatment rules. The treatment rule estimators can accommodate a variety of budget constraints of interest including settings with uncertain and or heterogeneous costs, quantity constraints, and settings where costs are not realized at the time of treatment. Another benefit is that the proposed rules can take advantage of well developed Bayesian estimation machinery. Lastly, the models were shown to be competitive against state-of-the-art alternatives in a simulation study and an empirical illustration was examined.

There are a number of considerations for future work. It would be of interest to determine if different prior choices, for example a sparsity-inducing prior or a normal prior with a different form for the covariance matrix than that considered here, would facilitate bounds of the type in Section 1.4.2 or yield modeling suggestions for higher dimensional feature spaces. Rather than using the Gibbs posterior to form treatment rules, it could also prove fruitful to approximate the Gibbs posterior with alternative distribution such as a normal distribution. So-called variational approximations of Gibbs posteriors for general PAC-Bayesian approaches are considered in Alquier et al. (2016). This could yield greater flexibility in terms of functional form constraints, beyond control over the variables included in the treatment rules featured here. It would also be of interest to incorporate estimated propensity scores into the PAC-Bayesian framework here and to explore how this impacts rates of convergence. Lastly, the analysis for balancing the primary welfare or regret objective against that of a secondary cost objective can be generalized to settings beyond the welfare-based potential outcomes framework. Balancing a secondary objective of concern could also be of interest in classification or regression settings.

Chapter 1 contains material being prepared for submission for academic publication. The dissertation author is the sole author of this material.

Table 1.1. Simulation welfare gains for models at different cost levels

Cost	PB-G	PB-MV	PB-B	R-NP	Ignore Cost
			$a=1$		
0.1	0.27	0.28	0.28	0.28	0.11
0.3	0.56	0.57	0.58	0.57	0.34
0.6	0.90	0.91	0.92	0.92	0.70
0.9	1.20	1.22	1.23	1.23	1.04
1.2	1.46	1.48	1.49	1.49	1.34
1.5	1.66	1.68	1.69	1.70	1.57
1.8	1.79	1.81	1.82	1.82	1.74
			$a=2$		
0.1	0.46	0.47	0.47	0.45	0.10
0.3	0.71	0.72	0.73	0.70	0.31
0.6	1.01	1.01	1.03	1.01	0.63
0.9	1.28	1.29	1.30	1.29	0.96
1.2	1.51	1.53	1.54	1.54	1.26
1.5	1.68	1.70	1.71	1.72	1.52
1.8	1.79	1.81	1.82	1.83	1.71
			$a=4$		
0.1	0.60	0.61	0.61	0.57	0.10
0.3	0.80	0.80	0.82	0.78	0.30
0.6	1.07	1.08	1.09	1.07	0.61
0.9	1.33	1.34	1.35	1.34	0.92
1.2	1.55	1.56	1.58	1.57	1.24
1.5	1.70	1.72	1.74	1.74	1.51
1.8	1.80	1.82	1.84	1.83	1.70

Figures for $a = 1$

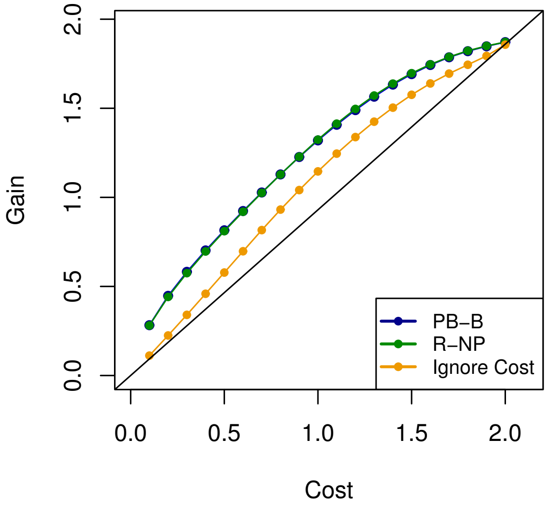


Figure 1.5. Cost curves when all methods utilize batch implementation for the DGP featuring $a = 1$.

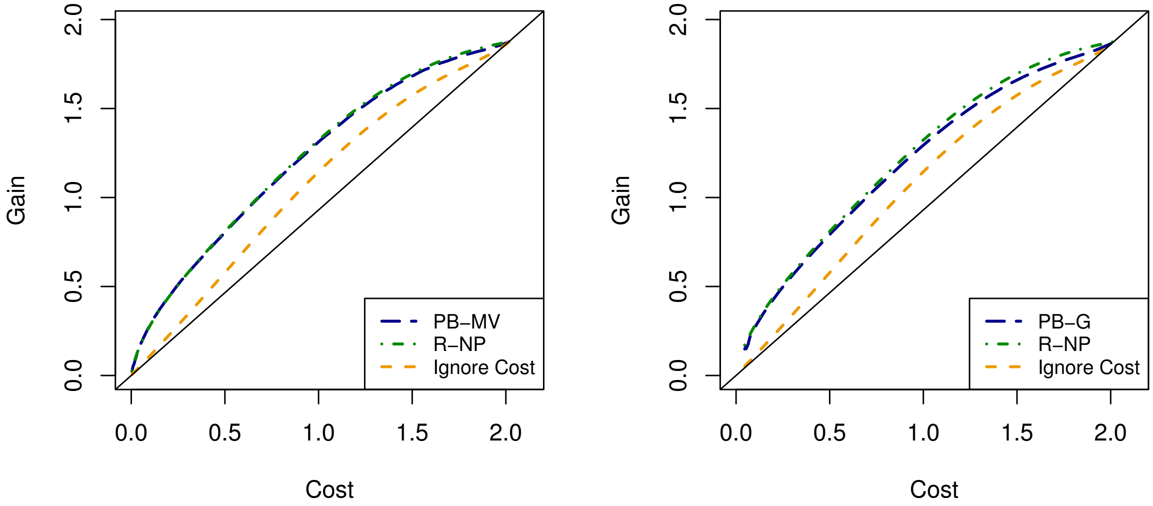


Figure 1.6. With $a = 1$ in the DGP, cost curves for the PB-MV and PB-G methods, which do not feature batch implementation, compared with the batch-implemented R-NP and IC methods.

Figures for $a = 2$

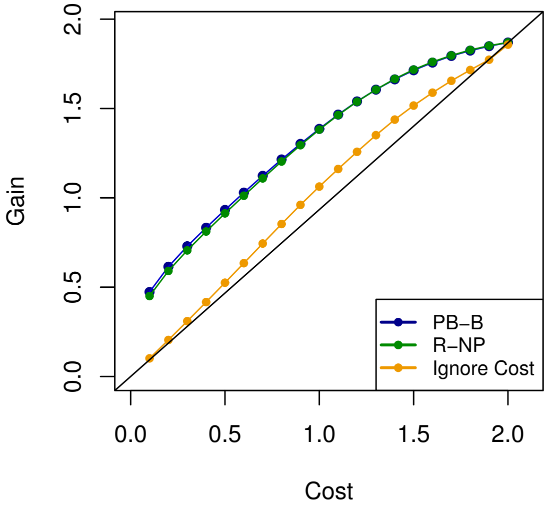


Figure 1.7. Cost curves when all methods utilize batch implementation for the DGP featuring $a = 2$.

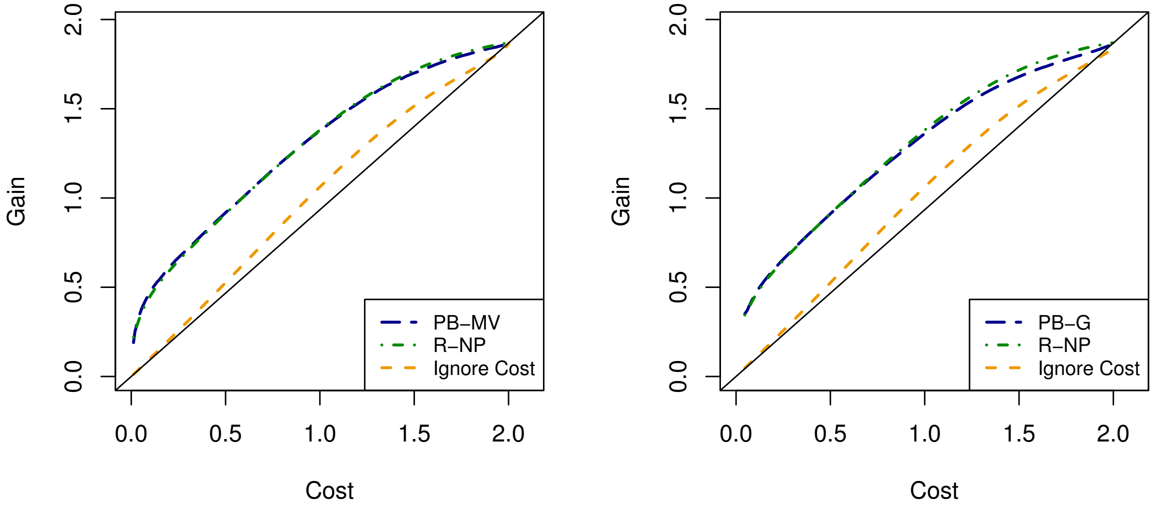


Figure 1.8. With $a = 2$ in the DGP, cost curves for the PB-MV and PB-G methods, which do not feature batch implementation, compared with the batch-implemented R-NP and IC methods.

Figures for $a = 4$

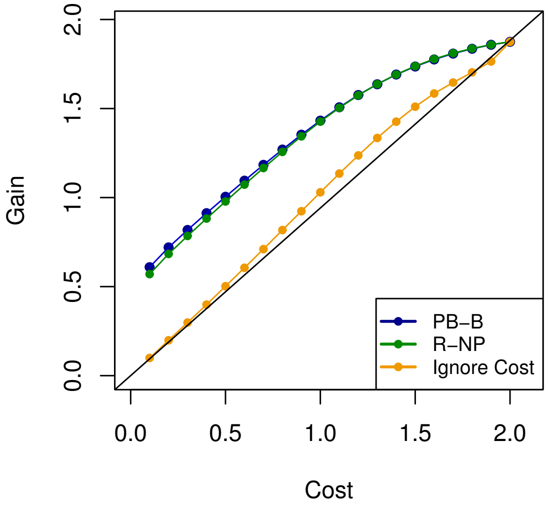


Figure 1.9. Cost curves when all methods utilize batch implementation for the DGP featuring $a = 4$.

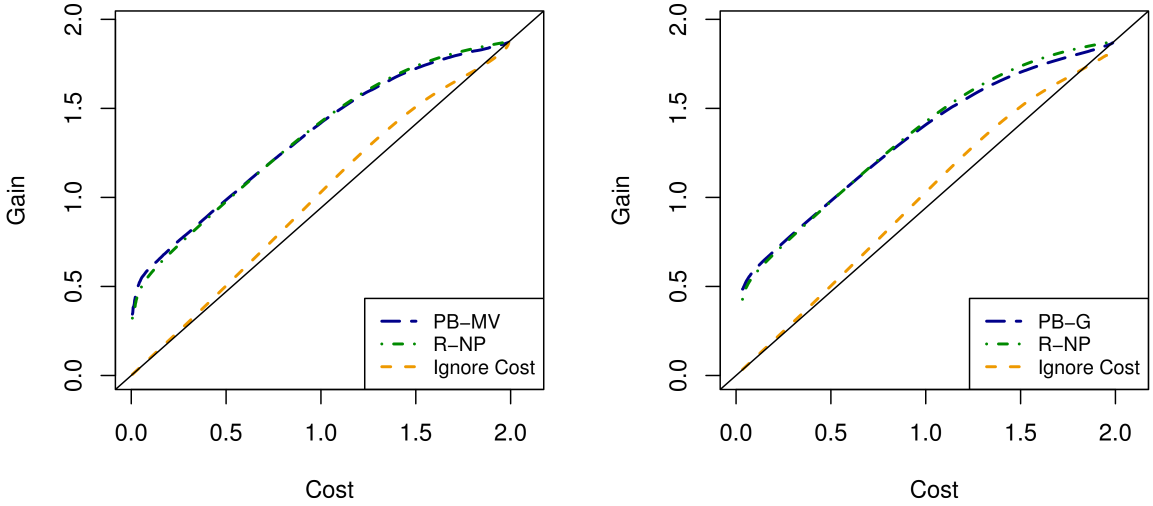


Figure 1.10. With $a = 4$ in the DGP, cost curves for the PB-MV and PB-G methods, which do not feature batch implementation, compared with the batch-implemented R-NP and IC methods.

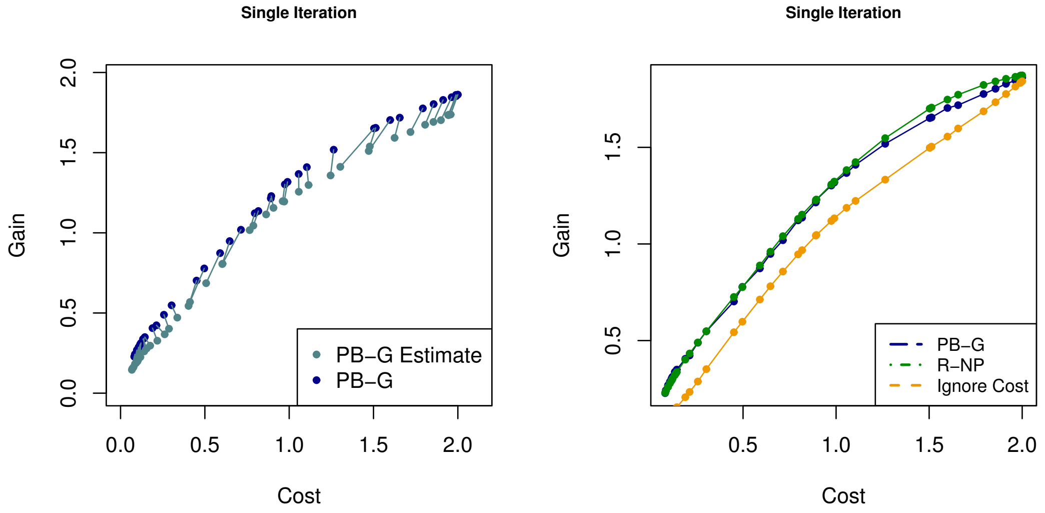


Figure 1.11. For the DGP with $a = 1$, the left-hand side plots the estimated and actual cost-gain pairs (one point for each u) for a single training sample iteration for the PB-G method. On the right-hand side the actual cost-gain pairs of PB-G models for various u values are plotted again, now compared with the actual cost-gain pairs associated with the R-NP and IC rules that produce the same target group cost. The points on the right are then interpolated to produce cost-gain curve estimates for a single iteration. These curves are averaged (vertically) over all simulation iterations to produce the right-hand side of Figure 1.6.

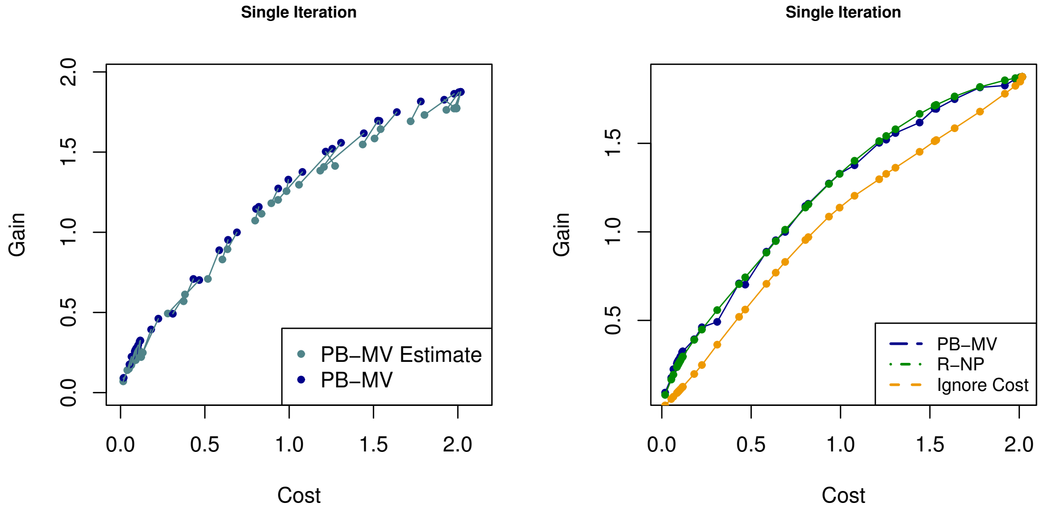


Figure 1.12. Illustrates a single training sample iteration for DGP1 when considering the PB-MV treatment model.

Appendices

1.A Appendix of Proofs for Chapter 1

1.A.1 Preliminaries and Adaptations From Earlier Literature to Our Setting

Here we consider preliminary properties to be utilized in subsequent analysis and recall results from the PAC-Bayesian literature that are also needed, sometimes with minor modifications. For the most part, proofs (and citations) are included for completeness even when a result is a fairly straightforward adaption.

Let $\mathscr{M}(\Theta)$ be the set of measurable functions on $(\Theta, \mathscr{B}_\theta)$ and let

$$\mathscr{M}_b^\pi(\Theta) = \left\{ A : A \in \mathscr{M}(\Theta) \text{ and } \int_\Theta \exp(A(\theta))\, d\pi(\theta) < \infty \right\},$$

which is a subset of $\mathscr{M}(\Theta)$ that has a finite exponential moment under π. We have the following lemma and corollary that will be utilized repeatedly in subsequent analysis. In particular they serve as a base in deriving Lemma 1.3.1 in Section 1.3.3.

Lemma 1.A.1 *For $\pi \in \mathscr{P}(\Theta)$ and $A \in \mathscr{M}(\Theta)$ such that $-A \in \mathscr{M}_b^\pi(\Theta)$, let $\rho_{A,\pi} \in \mathscr{P}_\pi(\Theta)$ be the probability measure on Θ with the Radon–Nikodym (RN) derivative with respect to π given by*

$$\frac{d\rho_{A,\pi}}{d\pi}(\theta) = \frac{\exp(-A(\theta))}{\int_\Theta \exp\left(-A\left(\tilde{\theta}\right)\right) d\pi\left(\tilde{\theta}\right)}.$$

Then for any probability measure $\rho \in \mathscr{P}_\pi(\Theta)$ we have

$$\log\left[\int_\Theta \exp\left(-A(\theta)\right) d\pi(\theta)\right] = -\left[\int_\Theta A(\theta)\, d\rho(\theta) + D_{\mathrm{KL}}(\rho,\pi)\right] + D_{\mathrm{KL}}\left(\rho, \rho_{A,\pi}\right). \quad (1.29)$$

Proof of Lemma 1.A.1. By definition,

$$\begin{aligned}
&D_{\mathrm{KL}}\left(\rho,\rho_{A,\pi}\right)\\
&= \int_\Theta \log\left[\frac{d\rho}{d\rho_{A,\pi}}(\theta)\right] d\rho(\theta)\\
&= \int_\Theta \log\left\{\frac{d\rho}{d\pi}(\theta)\left[\frac{d\rho_{A,\pi}}{d\pi}(\theta)\right]^{-1}\right\} d\rho(\theta)\\
&= \int_\Theta \left[\log\frac{d\rho}{d\pi}(\theta) - \log\frac{\exp\left(-A(\theta)\right)}{\int_\Theta \exp\left(-A\left(\tilde{\theta}\right)\right) d\pi\left(\tilde{\theta}\right)}\right] d\rho(\theta)\\
&= \int_\Theta A(\theta)\, d\rho(\theta) + \int_\Theta \log\left[\int_\Theta \exp\left(-A\left(\tilde{\theta}\right)\right) d\pi\left(\tilde{\theta}\right)\right] d\rho(\theta) + \int_\Theta \left[\log\frac{d\rho}{d\pi}(\theta)\right] d\rho(\theta)\\
&= \int_\Theta A(\theta)\, d\rho(\theta) + \log\left[\int_\Theta \exp\left(-A(\theta)\right) d\pi(\theta)\right] + \int_\Theta \left[\log\frac{d\rho}{d\pi}(\theta)\right] d\rho(\theta)\\
&= \int_\Theta A(\theta)\, d\rho(\theta) + \log\left[\int_\Theta \exp\left(-A(\theta)\right) d\pi(\theta)\right] + D_{\mathrm{KL}}(\rho,\pi).
\end{aligned}$$

Hence,

$$\log\left[\int_\Theta \exp\left(-A(\theta)\right) d\pi(\theta)\right] = -\left[\int_\Theta A(\theta)\, d\rho(\theta) + D_{\mathrm{KL}}(\rho,\pi)\right] + D_{\mathrm{KL}}\left(\rho, \rho_{A,\pi}\right).$$

■

Corollary 1.A.1 *(a) Let $\lambda > 0$, $\pi \in \mathscr{P}(\Theta)$, and let $A \in \mathscr{M}(\Theta)$ be such that $-\lambda A \in \mathscr{M}_b^\pi(\Theta)$. Then*

$$\rho_{\lambda A,\pi} = \underset{\rho \in \mathscr{P}_\pi(\Theta)}{\arg\min}\left[\int_\Theta A(\theta)\, d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho,\pi)\right],$$

and

$$\min_{\rho\in\mathscr{P}_\pi(\Theta)}\left[\int_\Theta A(\theta)\,d\rho(\theta)+\frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right]=-\frac{1}{\lambda}\log\left[\int_\Theta\exp(-\lambda A(\theta))\,d\pi(\theta)\right].$$

(b) For any $\mathscr{A}(\cdot)\in\mathscr{M}_b^\pi(\Theta)$, $\pi\in\mathscr{P}(\Theta)$, $\rho\in\mathscr{P}_\pi(\Theta)$,

$$\int_\Theta\mathscr{A}(\theta)\,d\rho(\theta)\le\log\left[\int_\Theta\exp(\mathscr{A}(\theta))\,d\pi(\theta)\right]+D_{\mathrm{KL}}(\rho,\pi).$$

Proof of Corollary 1.A.1. Part (a). Note $\rho_{\lambda A,\pi}=\arg\min_{\rho\in\mathscr{P}_\pi(\Theta)}D_{\mathrm{KL}}(\rho,\rho_{\lambda A,\pi})$ as $D_{\mathrm{KL}}(\rho,\pi)\ge$ 0 with equality if and only if $\rho=\pi$ π-almost surely. Replacing A with λA in Lemma 1.A.1 and noting that the left-hand-side of (1.29) does not vary with ρ we have

$$\begin{aligned}\rho_{\lambda A,\pi}&=\operatorname*{arg\,min}_{\rho\in\mathscr{P}_\pi(\Theta)}\left[D_{\mathrm{KL}}\left(\rho,\rho_{\lambda A,\pi}\right)\right]\\&=\operatorname*{arg\,min}_{\rho\in\mathscr{P}_\pi(\Theta)}\left[\int_\Theta\lambda A(\theta)d\rho(\theta)+D_{\mathrm{KL}}(\rho,\pi)\right]\\&=\operatorname*{arg\,min}_{\rho\in\mathscr{P}_\pi(\Theta)}\left[\int_\Theta A(\theta)d\rho(\theta)+\frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right].\end{aligned}$$

By equation (1.29) we then have

$$\begin{aligned}\min_{\rho\in\mathscr{P}(\Theta)}\left[\int_\Theta\lambda A(\theta)\,d\rho(\theta)+D_{\mathrm{KL}}(\rho,\pi)\right]&=\int_\Theta\lambda A(\theta)\,d\rho_{\lambda A,\pi}(\theta)+D_{\mathrm{KL}}\left(\rho_{\lambda A,\pi},\pi\right)\\&=-\log\left[\int_\Theta\exp(-\lambda A(\theta))\,d\pi(\theta)\right].\end{aligned}$$

This is equivalent to the second statement in part (a).

Part (b) Taking $A=-\mathscr{A}$ in Lemma 1.A.1, we obtain that for any probability measure $\rho\in\mathscr{P}_\pi(\Theta)$,

$$\log\left[\int_\Theta\exp(\mathscr{A}(\theta))\,d\pi(\theta)\right]=\left[\int_\Theta\mathscr{A}(\theta)\,d\rho(\theta)-D_{\mathrm{KL}}(\rho,\pi)\right]+D_{\mathrm{KL}}\left(\rho,\rho_{-A,\pi}\right).\quad(1.30)$$

Note that $D_{\mathrm{KL}}\left(\rho,\rho_{-A,\pi}\right)\geq 0$. It follows that

$$\begin{aligned}\log\left[\int_{\Theta}\exp\left(\mathscr{A}\left(\theta\right)\right)d\pi\left(\theta\right)\right] &= \left[\int_{\Theta}\mathscr{A}\left(\theta\right)d\rho\left(\theta\right)-D_{\mathrm{KL}}\left(\rho,\pi\right)\right]+D_{\mathrm{KL}}\left(\rho,\rho_{-A,\pi}\right)\\ &\geq\left[\int_{\Theta}\mathscr{A}\left(\theta\right)d\rho\left(\theta\right)-D_{\mathrm{KL}}\left(\rho,\pi\right)\right].\end{aligned}$$

This implies that

$$\int_{\Theta}\mathscr{A}\left(\theta\right)d\rho\left(\theta\right)\leq D_{\mathrm{KL}}\left(\rho,\pi\right)+\log\left[\int_{\Theta}\exp\left(\mathscr{A}\left(\theta\right)\right)d\pi\left(\theta\right)\right].$$

■

The following Theorem helps to produce PAC-Bayesian generalization bounds in our setting similar to counterparts in the classification literature. In particular, it essentially the same as Theorem 18 in Germain et al. (2015) with the loss function altered to the structure our setting; it is also similar to Theorem 4.1 in Alquier et al. (2016). The proof follows similar steps to those in Germain et al. (2015) and Alquier et al. (2016). We note that the proof applies to more general sample spaces, not just those following Assumption 1.3.1. We follow the current formulation to avoid additional exposition/notation.

Theorem 1.A.1 *Let Assumptions 1.3.1 and 1.3.2 (i) hold and let $\pi\in\mathscr{P}(\Theta)$. Let $\ell(Z,\theta):\mathscr{Z}\times\Theta\to\mathscr{R}$ denote a measurable loss function with range $\mathscr{R}\subseteq\mathbb{R}$. Define*

$$L(\theta)=E_P\left[\ell(Z,\theta)\right],\ L_n(\theta)=\frac{1}{n}\sum_{i=1}^{n}\ell(Z_i,\theta),$$

and, for $\rho\in\mathscr{P}_\pi(\Theta)$,

$$L\left(f_{G,\rho}\right)=\int_{\Theta}L(\theta)d\rho(\theta),\ L_n\left(f_{G,\rho}\right)=\int_{\Theta}L_n(\theta)d\rho(\theta).$$

Let $D : \mathscr{R} \times \mathscr{R} \to \mathbb{R}$ be any convex function and let $\lambda > 0$. Suppose

$$E_{P^n}\left[\int_{\Theta} \exp\left(\lambda D\left[L_n(\theta), L(\theta)\right]\right) d\pi(\theta)\right] \leq \exp\left(f(\lambda, n)\right), \tag{1.31}$$

where $f(\lambda, n) < \infty$ and may depend on λ and n. Then for any $\varepsilon \in (0,1]$ it holds with probability at least $1-\varepsilon$ that, simultaneously for all $\rho \in \mathscr{P}_\pi(\Theta)$,

$$D\left[L_n\left(f_{G,\rho}\right), L\left(f_{G,\rho}\right)\right] \leq \frac{f(\lambda, n) + \log\left(\frac{1}{\varepsilon}\right) + D_{\text{KL}}(\rho, \pi)}{\lambda}.$$

Proof of Theorem 1.A.1. (1.31) implies that

$$\int_{\Theta} \exp\left(\lambda D\left[L_n(\theta), L(\theta)\right]\right) d\pi(\theta) < \infty,$$

holds almost surely. Therefore, applying Corollary 1.A.1 (b) with $\mathscr{A}(\theta) = \lambda D[L_n(\theta), L(\theta)]$, the event

$$\left\{\int_{\Theta} \lambda D[L_n(\theta), L(\theta)] d\rho(\theta) \right.$$
$$\left. \leq \log\left[\int_{\Theta} \exp\left(\lambda D\left[L_n(\theta), L(\theta)\right]\right) d\pi(\theta)\right] + D_{\text{KL}}(\rho, \pi) \text{ for all } \rho \in \mathscr{P}_\pi(\Theta) \text{ simultaneously}\right\},$$

occurs with probability one. Applying Jensen's inequality to the object on the left-hand-side of the inequality in this event, we have

$$\begin{aligned} P^n & \left\{\lambda D[L_n\left(f_{G,\rho}\right), L\left(f_{G,\rho}\right)] \right. \\ & \qquad \left. \leq \log\left[\int_{\Theta} \exp\left(\lambda D\left[L_n(\theta), L(\theta)\right]\right) d\pi(\theta)\right] + D_{\text{KL}}(\rho, \pi) \text{ for all } \rho \in \mathscr{P}_\pi(\Theta) \text{ simultaneously}\right\}, \\ & = 1 \end{aligned} \tag{1.32}$$

By Markov's inequality and then applying (1.31),

$$
\begin{aligned}
&P^n\left\{\int_\Theta \exp\left(\lambda D\left[L_n(\theta),L(\theta)\right]\right)d\pi(\theta) > \exp\left[f(\lambda,n)+\log\left(\frac{1}{\varepsilon}\right)\right]\right\}\\
&\leq \frac{E_{P^n}\left[\int_\Theta \exp\left(\lambda D\left[L_n(\theta),L(\theta)\right]\right)d\pi(\theta)\right]}{\exp\left[f(\lambda,n)+\log\left(\frac{1}{\varepsilon}\right)\right]}\\
&\leq \varepsilon.
\end{aligned}
$$

Therefore,

$$
P^n\left\{\log\left[\int_\Theta \exp\left(\lambda D\left[L_n(\theta),L(\theta)\right]\right)d\pi(\theta)\right] \leq f(\lambda,n)+\log\left(\frac{1}{\varepsilon}\right)\right\} \geq 1-\varepsilon
$$

Note that this high probability bound does not involve ρ. Combining it with (1.32), we have

$$
\begin{aligned}
&P^n\left\{D[L_n\left(f_{G,\rho}\right),L\left(f_{G,\rho}\right)] \leq \frac{f(\lambda,n)+\log\left(\frac{1}{\varepsilon}\right)+D_{\text{KL}}(\rho,\pi)}{\lambda} \text{ for all } \rho\in\mathscr{P}_\pi(\Theta) \text{ simultaneously}\right\}\\
&\geq 1-\varepsilon
\end{aligned}
$$

■

The following lemma will be combined with Theorem 1.A.1 to produce Theorem 1.A.2 below. The lemma yields a key step in adapting PAC-Bayesian bounds from the 0/1-loss classification literature to more general settings, a procedure utilized in Maurer (2004) and Germain et al. (2015). For us it will allow us to follow those author's adaption of a well known PAC-Bayesian bound, appearing, for example, in Seeger (2002), to more general settings. This then serves as a key input for producing Lemma 1.A.3 following the analysis of Lever et al. (2010).

Lemma 1.A.2 *Let X be any random variable taking values in $[0,1]$ with $EX=\mu$. Denote $\mathbf{X}=(X_1,\ldots,X_n)$ where $X_1,\ldots,X_n$ are iid realizations of X. Let $\mathbf{X}'=(X_1',\ldots,X_n')$ where $X_1',\ldots,X_n'$ are iid realizations of a Bernoulli random variable X' with probability of success μ. If $f:[0,1]^n\to\mathbb{R}$*

is convex, then

$$E\left[f\left(\mathbf{X}\right)\right] \leq E\left[f\left(\mathbf{X}'\right)\right]$$

Proof of Lemma 1.A.2. This lemma is due to Maurer (2004). Another proof with more details is given in Germain et al. (2015); see Lemmas 51 and 52 there. For intuition, we can regard $\mathbf{X}'$ as a mean-preserving spread of $\mathbf{X}$ and $-f$ as the utility function. Then the lemma says that $\mathbf{X}$ is preferred by an expected utility maximizer having concave utility $-f\left(\cdot\right)$. ■

Now we use Lemma 1.A.2 combined with Theorem 1.A.1 to produce Theorem 1.A.2 below, which is a variant of a well known bound appearing in Seeger (2002). To do this, we follow the analysis in Germain et al. (2015) to verify the bound for our setting. The proof closely follows that in Germain et al. (2015). Theorem 20 in Germain et al. (2015), for example, is a very similar and can apply to a variety of settings. The only difference here is that the structure of what plays the role of a loss function is stated differently in Theorem 1.A.1.

The following notation is used in the next theorem. We let

$$\mathrm{kl}(a,b) = a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}, \tag{1.33}$$

and adopt the convention that $0\log 0 = 0$, $a\log\frac{a}{0} = \infty$ if $a > 0$ and $0\log\frac{0}{0} = 0$. Note that $\mathrm{kl}(a,b)$ is the KL-divergence between two Bernoulli random variables with success probabilities a and b.

Theorem 1.A.2 *Set any prior $\pi \in \mathscr{P}(\theta)$ and $\varepsilon \in (0,1]$. Let Assumption 1.3.1, 1.3.2, and 1.3.3 hold. Let $\ell(Z,\theta) : \mathscr{Z} \times \Theta \to [0,1]$ denote a measurable loss function with range $[0,1]$ (equipped with the standard Borel sigma field). Define*

$$L(\theta) = E_P\left[\ell(Z,\theta)\right],\ \ L_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\ell(Z_i,\theta),$$

and, for $\rho \in \mathscr{P}_\pi(\Theta)$*,*

$$L\left(f_{G,\rho}\right) = \int_\Theta L(\theta) d\rho(\theta),\ \ L_n\left(f_{G,\rho}\right) = \int_\Theta L_n(\theta) d\rho(\theta).$$

(a). With probability at least $1-\varepsilon$*, for all posteriors* $\rho \in \mathscr{P}_\pi(\Theta)$ *simultaneously it holds that*

$$\mathrm{kl}\left(L_n\left(f_{G,\rho}\right), L\left(f_{G,\rho}\right)\right) \leq \frac{1}{n}\left[D_{\mathrm{KL}}(\rho,\pi) + \log\left(2\sqrt{n}\right) + \log\frac{1}{\varepsilon}\right].$$

(b). With probability at least $1-\varepsilon$*, for all posteriors* $\rho \in \mathscr{P}_\pi(\Theta)$ *simultaneously it holds that*

$$\left(L_n\left(f_{G,\rho}\right) - L\left(f_{G,\rho}\right)\right)^2 \leq \frac{1}{2n}\left[D_{\mathrm{KL}}(\rho,\pi) + \log\left(2\sqrt{n}\right) + \log\frac{1}{\varepsilon}\right].$$

Proof of Theorem 1.A.2. Part (a) Given the adaptation of Theorem 1.A.1 to our setting, the proof follows that of Lemma 19 in Germain et al. (2015) or Theorem 1 in Maurer (2004). We will apply Theorem 1.A.1 with

$$D(a,b) = \frac{n}{\lambda}\mathrm{kl}(a,b).$$

That $\mathrm{kl}(\cdot,\cdot)$ is convex follows from Theorem 2.7.2 of Cover and Thomas (2006). We must verify that the condition in (1.31) holds with $f(\lambda,n) = \log(2\sqrt{n})$. We will show that for any $\theta \in \Theta$,

$$E_{P^n}\left\{\exp\left[n\mathrm{kl}\left(L_n(\theta), L(\theta)\right)\right]\right\} \leq \sum_{k=0}^{n}\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} \equiv \xi(n). \tag{1.34}$$

It can be shown (c.f. Lemma 19 in Germain et al. (2015) and the references therein) that $\sqrt{n} \leq \xi(n) \leq 2\sqrt{n}$. Then, by Assumption 1.3.3, we can reverse the order of integration on the object on the left hand side of condition 1.31, so that (1.34) yields that (1.31) holds with $f(\lambda,n) = \log(2\sqrt{n})$. All that remains is to prove (1.34).

Let $\theta \in \Theta$. First note that in edge cases where $L(\theta) = 0$ or $L(\theta) = 1$, we then have with probability one that $L_n(\theta) = 0$ or $L_n(\theta) = 1$, respectively, in which case $\mathrm{kl}(L_n(\theta), L(\theta)) = 0$

and (1.34) holds. Now consider any θ such that $L(\theta) \in (0,1)$. Note that

$$\exp\{\lambda D(L_n(\theta), L(\theta))\} = \exp\left\{ n \cdot \mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n} \ell(Z_i, \theta), L(\theta) \right) \right\}$$

is a convex function of $\mathbf{X} = (\ell(Z_1,\theta), \ldots, \ell(Z_n,\theta))$. Then, by Lemma 1.A.2,

$$E_{P^n}\{\exp\{\lambda D(L_n(\theta), L(\theta))\}\} \le E \exp\left\{ n \cdot \mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^{n} X_i', L(\theta) \right) \right\}, \tag{1.35}$$

where $X_1', \ldots, X_n'$ are iid Bernoulli random variables with success probability $L(\theta)$ and the expectation on the right is taken with respect to their joint distribution. Denoting $X' = \sum_{i=1}^{n} X_i'$, we have

$$\begin{aligned}
& E \exp\left\{ n \cdot \mathrm{kl}\left(\frac{1}{n} X', L(\theta) \right) \right\} \\
&\quad = E \left(\frac{\frac{1}{n}X'}{L(\theta)} \right)^{X'} \left(\frac{1 - \frac{1}{n}X'}{1 - L(\theta)} \right)^{n-X'} \\
&\quad = \sum_{k=0}^{n} \Pr\left(X' = k\right) \left(\frac{\frac{k}{n}}{L(\theta)} \right)^{k} \left(\frac{1 - \frac{k}{n}}{1 - L(\theta)} \right)^{n-k} \\
&\quad = \sum_{k=0}^{n} \binom{n}{k} (L(\theta))^k (1 - L(\theta))^{n-k} \left(\frac{\frac{k}{n}}{L(\theta)} \right)^{k} \left(\frac{1 - \frac{k}{n}}{1 - L(\theta)} \right)^{n-k} \\
&\quad = \sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n} \right)^{k} \left(1 - \frac{k}{n} \right)^{n-k} = \xi(n)
\end{aligned} \tag{1.36}$$

Therefore (1.34) holds for any $\theta \in \Theta$, completing the proof.

Part (b). Part (b) follows from part (a) with an application of Pinsker's inequality,

$$2(a-b)^2 \le \mathrm{kl}(a,b) \tag{1.37}$$

■

The following lemma adapts Lemma 2 of Lever et al. (2010) to our setting, it will aid in removing a D_{KL} term from several bounds in Section 1.4.

Lemma 1.A.3 *Let $\hat{\rho}_{\lambda,u}$ and $\rho^*_{\lambda,u}$ be as in Definition 1.3.2 with $\pi \in \mathscr{P}(\Theta)$, $\lambda > 0$, and $u \geq 0$. Let Assumptions 1.3.1, 1.3.2, and 1.3.3 hold and let $\varepsilon \in (0,1]$. With probability at least $1-\varepsilon$ it holds that*

$$D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,u}, \rho^*_{\lambda,u}\right) \leq \frac{\lambda\sqrt{2}\,(M_y + uM_c)}{\kappa\sqrt{n}}\sqrt{\log\left(\frac{2\sqrt{n}}{\varepsilon}\right)} + \frac{\lambda^2\,(M_y+uM_c)^2}{2n\kappa^2}.$$

Proof of Lemma 1.A.3. The proof follows that of Lemma 2 in Lever et al. (2010), with some minor adjustments, which are straightforward with Theorem 1.A.2 taking the place of Seeger's (c.f. Seeger (2002)) bound in the setting of Lever et al. (2010). To lighten the exposition, we will write

$$M(\theta;u) = R(\theta) + uK(\theta) \text{ and } M_n(\theta;u) = R_n(\theta) + uK_n(\theta)$$

when writing the RN deriviatives of $\hat{\rho}_{\lambda,u}$ and $\rho^*_{\lambda,u}$ with respect to π and related objects. Note that for any $\theta \in \Theta$ we have $M_n(\theta;u) \in [-(M_y+uM_c)/2\kappa, (M_y+uM_c)/2\kappa]$ by Assumption 1.3.1 (iii) and (iv).

First, observe that

$$
\begin{aligned}
& D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,u}, \rho^*_{\lambda,u}\right) \\
&= \int_\Theta \log\left[\left(\frac{d\hat{\rho}_{\lambda,u}}{d\pi}(\theta)\right)\left(\frac{d\pi}{d\rho^*_{\lambda,u}}(\theta)\right)\right] d\hat{\rho}_{\lambda,u}(\theta) \\
&= \int_\Theta \left(\log\left[\frac{\exp\left(-\lambda M_n(\theta;u)\right)}{\exp\left(-\lambda M(\theta;u)\right)}\right] - \log\left[\frac{\int_\Theta \exp\left(-\lambda M_n(\theta,u)\right) d\pi(\theta)}{\int_\Theta \exp\left(-\lambda M(\theta;u)\right) d\pi(\theta)}\right]\right) d\hat{\rho}_{\lambda,u}(\theta) \\
&= \int_\Theta \log\left[\frac{\exp\left(-\lambda M_n(\theta;u)\right)}{\exp\left(-\lambda M(\theta;u)\right)}\right] d\hat{\rho}_{\lambda,u}(\theta) \\
&\qquad - \log\left[\frac{\int_\Theta \exp\left(-\lambda\left[M_n(\theta;u) + M(\theta;u) - M(\theta;u)\right]\right) d\pi(\theta)}{\int_\Theta \exp\left(-\lambda M(\theta;u)\right) d\pi(\theta)}\right] \\
&= \lambda \int_\Theta M(\theta;u) - M_n(\theta;u) d\hat{\rho}_{\lambda,u}(\theta) - \log\left[\int_\Theta \exp\left(\lambda\left[M(\theta;u) - M_n(\theta;u)\right]\right) d\rho^*_{\lambda,u}\right] \\
&\leq \lambda\left[\int_\Theta M(\theta;u) - M_n(\theta;u) d\hat{\rho}_{\lambda,u}(\theta) - \int_\Theta M(\theta;u) - M_n(\theta;u) d\rho^*_{\lambda,u}\right], \qquad (1.38)
\end{aligned}
$$

where the last inequality follows from Jensen's inequality.

Next we utilize an Theorem 1.A.2 (b). For the setting there, let

$$
\ell(Z,\theta) = \left(\ell_y(Z,\theta) + u\ell_c(Z,\theta) + \frac{M_y + uM_c}{2\kappa}\right)\left(\frac{\kappa}{M_y + uM_c}\right)
$$

where

$$
\ell_y(Z,\theta) = \left(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\right)\left(f^*(X) - f_\theta(X)\right), \qquad (1.39)
$$

$$
\ell_c(Z,\theta) = \left(\frac{CD}{e(X)} - \frac{C(1-D)}{1-e(X)}\right) f_\theta(X), \qquad (1.40)
$$

and f^* is as in (1.7).

Note then that, by Assumption (1.3.1) (iii) and (iv), for all $\theta \in \Theta$, we have $\ell(Z,\theta) \in [0,1]$

almost surely. Additionally, we have

$$\begin{aligned} L(\theta) = E_P[\ell(Z,\theta)] &= \left(R(\theta) + uK(\theta) + \frac{M_y + uM_c}{2\kappa}\right)\left(\frac{\kappa}{M_y + uM_c}\right) \\ &= \left(M(\theta;u) + \frac{M_y + uM_c}{2\kappa}\right)\left(\frac{\kappa}{M_y + uM_c}\right) \end{aligned}$$

and

$$\begin{aligned} L_n(\theta) &= \left(R_n(\theta) + uK_n(\theta) + \frac{M_y + uM_c}{2\kappa}\right)\left(\frac{\kappa}{M_y + uM_c}\right) \\ &= \left(M_n(\theta;u) + \frac{M_y + uM_c}{2\kappa}\right)\left(\frac{\kappa}{M_y + uM_c}\right). \end{aligned}$$

Given the above setting, we will apply Theorem 1.A.2 (b). Note that in Theorem 1.A.2, the prior π does not have to be the same as that used in the definition of $\hat{\rho}_{\lambda,u}$ and $\rho^*_{\lambda,u}$, provided that the posteriors of interest are still absolutely continuous with respect to the prior. Rather that utilizing the theorem with the π associated with $\hat{\rho}_{\lambda,u}$ and $\rho^*_{\lambda,u}$, we instead use $\rho^*_{\lambda,u}$ as the prior. Note this prior choice satisfies Assumption 1.3.3, i.e. it does not depend on the sample. Applying Theorem 1.A.2 (b) and taking the square root of each side in the high probability bound there, utilizing posteriors $\rho = \hat{\rho}_{\lambda,u}$ and $\rho = \rho^*_{\lambda,u}$, with probability at least $1-\varepsilon$ it holds simultaneously that

$$\begin{aligned} \int_\Theta L(\theta) - L_n(\theta) d\hat{\rho}_{\lambda,u}(\theta) &\le \frac{1}{\sqrt{2n}}\sqrt{D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,u}, \rho^*_{\lambda,u}\right) + \log\left(\frac{2\sqrt{n}}{\varepsilon}\right)}, \\ -\left(\int_\Theta L(\theta) - L_n(\theta) d\rho^*_{\lambda,u}(\theta)\right) &\le \frac{1}{\sqrt{2n}}\sqrt{\log\left(\frac{2\sqrt{n}}{\varepsilon}\right)}. \end{aligned}$$

In terms of $M(\theta;u)$ and $M_n(\theta;u)$, this reads: with probability at least $1-\varepsilon$, the following events holds simultaneously

$$\int_{\Theta} M(\theta) - M_n(\theta) d\hat{\rho}_{\lambda,u}(\theta) \leq \frac{M_y + uM_c}{\kappa\sqrt{2n}} \sqrt{D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,u}, \rho^*_{\lambda,u}\right) + \log\left(\frac{2\sqrt{n}}{\varepsilon}\right)},$$
$$-\left(\int_{\Theta} M(\theta) - M_n(\theta) d\rho^*_{\lambda,u}(\theta)\right) \leq \frac{M_y + uM_c}{\kappa\sqrt{2n}} \sqrt{\log\left(\frac{2\sqrt{n}}{\varepsilon}\right)}.$$

Applying the above two inequalities to (1.38), we obtain

$$\begin{aligned} & D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,u}, \rho^*_{\lambda,u}\right) \\ & \leq \frac{\lambda\left(M_y + uM_c\right)}{\kappa\sqrt{2n}} \sqrt{D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,u}, \rho^*_{\lambda,u}\right) + \log\left(\frac{2\sqrt{n}}{\varepsilon}\right)} + \frac{\lambda\left(M_y + uM_c\right)}{\kappa\sqrt{2n}} \sqrt{\log\left(\frac{2\sqrt{n}}{\varepsilon}\right)} \end{aligned}$$

Straightforward algebraic manipulations of the above produce that

$$\begin{aligned} & \left(D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,u}, \rho^*_{\lambda,u}\right)\right)^2 - \frac{2\lambda\left(M_y + uM_c\right)}{\kappa\sqrt{2n}} \sqrt{\log\left(\frac{2\sqrt{n}}{\varepsilon}\right)} D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,u}, \rho^*_{\lambda,u}\right) + \frac{\lambda^2\left(M_y + uM_c\right)^2}{2n\kappa^2} \log\left(\frac{2\sqrt{n}}{\varepsilon}\right) \\ & \leq \frac{\lambda^2\left(M_y + uM_c\right)^2}{2n\kappa^2} D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,u}, \rho^*_{\lambda,u}\right) + \frac{\lambda^2\left(M_y + uM_c\right)^2}{2n\kappa^2} \log\left(\frac{2\sqrt{n}}{\varepsilon}\right). \end{aligned} \tag{1.41}$$

If

$$D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,u}, \rho^*_{\lambda,u}\right) \leq \frac{2\lambda\left(M_y + uM_c\right)}{\kappa\sqrt{2n}} \sqrt{\log\left(\frac{2\sqrt{n}}{\varepsilon}\right)},$$

the statement of the lemma holds. Otherwise, this and the fact that $D_{\mathrm{KL}}(\hat{\rho}_{\lambda,u}, \rho^*_{\lambda,u}) \geq 0$ imply that $D_{\mathrm{KL}}(\hat{\rho}_{\lambda,u}, \rho^*_{\lambda,u}) > 0$. Then, canceling out terms on either side of the inequality in (1.41) and dividing each side by $D_{\mathrm{KL}}(\hat{\rho}_{\lambda,u}, \rho^*_{\lambda,u})$ produces the statement of the lemma. ■

The remainder of the section contains straightforward lemmas that will be utilized in proofs for results in Section 1.4 and one more substantial result adapted from Freund et al. (2004) that will conclude this subsection.

Lemma 1.A.4 *Let Assumptions 1.3.1 and 1.3.2 hold. Let $\rho' \in \mathscr{P}(\Theta)$ be a (deterministic)* *probability that does not depend on the sample. Then*

$$P^n\left(\int_\Theta K_n(\theta)d\rho'(\theta) \leq \int_\Theta K(\theta)d\rho'(\theta) + \sqrt{\frac{M_c^2 \log(1/\varepsilon)}{2n\kappa^2}}\right) \geq 1-\varepsilon.$$

Proof of Lemma 1.A.4. Define the mapping

$$K(Z_1,\ldots,Z_n) = \int_\Theta K_n(\theta)d\rho'(\theta).$$

It is straightforward to check that, under Assumption 1.3.1 (iii), K satisfies the bounded differences property in Section 6.1 of Boucheron et al. (2013) with (in their notation) $c_i = M_c/(n\kappa)$ for $i = 1,\ldots,n$. It follows by McDiarmid's inequality (c.f. McDiarmid (1989)) that, for any $t \geq 0$,

$$\begin{aligned} &P^n\left(\int_\Theta K_n(\theta)d\rho'(\theta) - E_{P^n}\left[\int_\Theta K_n(\theta)d\rho'(\theta)\right] > t\right) \\ &= P^n\left(\int_\Theta K_n(\theta)d\rho'(\theta) - \int_\Theta K(\theta)d\rho'(\theta) > t\right) \leq \exp\left\{-\frac{2n\kappa^2 t^2}{M_c^2}\right\}. \end{aligned}$$

Substituting $t = \sqrt{M_c^2 \log(1/\varepsilon)/(2n\kappa^2)}$, for any $\varepsilon \in (0,1]$, this says

$$P^n\left(\int_\Theta K_n(\theta)d\rho'(\theta) - \int_\Theta K(\theta)d\rho'(\theta) > \sqrt{\frac{M_c^2 \log(1/\varepsilon)}{2n\kappa^2}}\right) \leq \varepsilon.$$

The result follows by taking the compliment and rearranging terms. ■

Lemma 1.A.5 *The KL divergence between $\rho : N(\mu_\rho, \Sigma_\rho)$ and $\pi : N(\mu_\pi, \Sigma_\pi)$ on $\mathbb{R}^q$, where μ_θ and μ_ρ are mean vectors and Σ_π and Σ_ρ are covariance matrices, is*

$$D_{\mathrm{KL}}(\rho,\pi) = \frac{1}{2}\left(\mu_\rho - \mu_\pi\right)'\Sigma_\pi^{-1}\left(\mu_\rho - \mu_\pi\right) + \frac{1}{2}\left[\operatorname{tr}\left(\Sigma_\rho\Sigma_\pi^{-1}\right) - q\right] - \frac{1}{2}\log\frac{\det\left(\Sigma_\rho\right)}{\det\left(\Sigma_\pi\right)}.$$

Proof of Lemma 1.A.5. By definition and via simple calculations, we have

$$
\begin{aligned}
& D_{\mathrm{KL}}(\rho,\pi) \\
&= -\frac{1}{2}E_{\theta\sim\rho}\left[\log\frac{\det(\Sigma_\rho)}{\det(\Sigma_\pi)} + (\theta-\mu_\rho)'\Sigma_\rho^{-1}(\theta-\mu_\rho) - (\theta-\mu_\pi)'\Sigma_\pi^{-1}(\theta-\mu_\pi)\right] \\
&= -\frac{1}{2}\log\frac{\det(\Sigma_\rho)}{\det(\Sigma_\pi)} - \frac{1}{2}\left[q - E_{\theta\sim\rho}(\theta-\mu_\rho+\mu_\rho-\mu_\pi)'\Sigma_\pi^{-1}(\theta-\mu_\rho+\mu_\rho-\mu_\pi)\right] \\
&= -\frac{1}{2}\log\frac{\det(\Sigma_\rho)}{\det(\Sigma_\pi)} - \frac{1}{2}\left[q - tr\left(\Sigma_\rho\Sigma_\pi^{-1}\right) - (\mu_\rho-\mu_\pi)'\Sigma_\pi^{-1}(\mu_\rho-\mu_\pi)\right] \\
&= \frac{1}{2}(\mu_\rho-\mu_\pi)'\Sigma_\pi^{-1}(\mu_\rho-\mu_\pi) + \frac{1}{2}\left[tr\left(\Sigma_\rho\Sigma_\pi^{-1}\right) - q\right] - \frac{1}{2}\log\frac{\det(\Sigma_\rho)}{\det(\Sigma_\pi)}.
\end{aligned}
$$

■

The last results needed for our analysis are stated in the two lemmas below. The first is a more elementary property used in proving the second, which is utilized during a step in the proof of Theorem 1.4.2 in Section 1.4. Both are close adaptions of analysis in Freund et al. (2004). After a translation of the problem via Corollary 1.A.1, we follow the method of proof there, adapting the analysis there in the 0/1 loss setting to ours with fairly straightforward modifications.

Lemma 1.A.6 *For $x = (x_1,\ldots,x_m) \in \mathbb{R}^m$, and with $\{a_i\}_{i=1}^m$ such that $a_i \geq 0$ for all $i = 1,\ldots,m$, the function*

$$x \mapsto -\log\left[\sum_{i=1}^m a_i \exp[x_i]\right]$$

is concave.

Proof of Lemma 1.A.6. Let $\alpha \in (0,1)$ and $x,y \in \mathbb{R}^m$. We will show that

$$K(x) = \log\left[\sum_{i=1}^m a_i \exp[x_i]\right]$$

is convex. Let $p = 1/\alpha$, $q = 1/(1-\alpha)$ and define $r_i = a_i^{1/p}\exp[\alpha x_i]$ and $s_i = a_i^{1/q}\exp[(1-\alpha)y_i]$.

As $1/p+1/q=1$, by Hölder's inequality,

$$\sum_{i=1}^{m} r_i s_i \leq \left(\sum_{i=1}^{m} r_i^p\right)^{1/p} \left(\sum_{i=1}^{m} s_i^q\right)^{1/q}.$$

Taking the logarithm of each side and plugging in the definitions of p, q, r_i and s_i, this is equivalent to

$$K(\alpha x+(1-\alpha)y) \leq \alpha K(x)+(1-\alpha)K(y),$$

completing the proof. ■

The following lemma combines pieces of Lemmas 1 and 2 of Freund et al. (2004) and translates those results for the 0/1-loss setting to a useful ingredient for ours.

Lemma 1.A.7 *Let $\hat{\rho}_{\lambda,u}$ and $\rho^*_{\lambda,u}$ be as in Definition 1.3.2 with $\pi \in \mathscr{P}(\Theta)$, $\lambda > 0$, and $u \geq 0$. Let assumptions 1.3.1, 1.3.2, and 1.3.3 hold. Then, for any $\varepsilon \in (0,1]$, it holds that*

$$\begin{aligned}
&P^n \Bigg\{ \int_\Theta R_n(\theta) d\hat{\rho}_{\lambda,u}(\theta) + u \int_\Theta K_n(\theta) d\hat{\rho}_{\lambda,u}(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,u},\pi\right) \leq \\
&\quad \int_\Theta R(\theta) d\rho^*_{\lambda,u}(\theta) + u \int_\Theta K(\theta) d\rho^*_{\lambda,u}(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho^*_{\lambda,u},\pi) + \sqrt{\frac{(M_y+uM_c)^2 \log(1/\varepsilon)}{2n\kappa^2}} \Bigg\} \\
&\quad \geq 1-\varepsilon.
\end{aligned}$$

Proof of Lemma 1.A.7. Define the mapping

$$K_u(Z_1,\ldots,Z_n) = \int_\Theta R_n(\theta) d\hat{\rho}_{\lambda,u}(\theta) + u \int_\Theta K_n(\theta) d\hat{\rho}_{\lambda,u}(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,u},\pi\right).$$

Note that by Corollary 1.A.1 (a), replacing $A(\theta)$ in the Corollary with $R(\theta)+uK_n(\theta)$,

$$K_u(Z_1,\ldots,Z_n) = -\frac{1}{\lambda} \log\left[\int_\Theta \exp\left[-\lambda\left(R_n(\theta)+uK_n(\theta)\right)\right] d\pi(\theta)\right]. \tag{1.42}$$

First we show that for any $\varepsilon \in (0,1]$ it holds that

$$P^n\left(K_u(Z_1,\dots,Z_n) > E_{P^n}\left[K_u(Z_1,\dots,Z_n)\right] + \sqrt{\frac{(M_y+uM_c)^2\log(1/\varepsilon)}{2n\kappa^2}}\right) \leq \varepsilon. \qquad (1.43)$$

To show this, for any $i \in \{1,\dots,n\}$, let $Z_i' \in \mathscr{Z}$ and let $(Z_1,\dots,Z_n) \in \mathscr{Z}^n$. Let $K_n(\theta)$ and $R_n(\theta)$ be computed utilizing $(Z_1,\dots,Z_{i-1},Z_i,Z_{i+1},\dots,Z_n)$ and let $K_n'(\theta)$ and $R_n'(\theta)$ be computed as $K_n(\theta)$ and $R_n(\theta)$ are, respectively, except utilizing the sample $(Z_1,\dots,Z_{i-1},Z_i',Z_{i+1},\dots,Z_n)$ instead of $(Z_1,\dots,Z_{i-1},Z_i,Z_{i+1},\dots,Z_n)$. Also, let $K_{n-i}(\theta)$ and R_{n-i} denote the computation of $K_n(\theta)$ and $R_n(\theta)$, respectively, except with the sample of size $n-1$ that drops observation Z_i. Then by construction $K_{n-i}(\theta) = K_{n-1}'(\theta)$ and $R_{n-i}(\theta) = R_{n-1}'(\theta)$. Under Assumptions 1.3.1 (iii) and (iv),

$$-\frac{M_y+uM_c}{2\kappa} \leq \ell_y(Z_i) + u\ell_c(Z_i,\theta) \leq \frac{M_y+uM_c}{2\kappa}$$

almost surely where ℓ_y and ℓ_c are defined in (1.39) and (1.40) and are summed over i in $R_n(\theta)$ and $K_n(\theta)$, respectively. It follows from (1.42) that,

$$\begin{aligned}
&|K_u(Z_1,\dots,Z_{i-1},Z_i,Z_{i+1},\dots,Z_n) - K_u(Z_1,\dots,Z_{i-1},Z_i',Z_{i+1},\dots,Z_n)| \\
&= \left| -\frac{1}{\lambda}\log\left[\frac{\int_\Theta \exp\left[-\lambda\left(R_n(\theta)+uK_n(\theta)\right)\right]d\pi(\theta)}{\int_\Theta \exp\left[-\lambda\left(R_n'(\theta)+uK_n'(\theta)\right)\right]d\pi(\theta)}\right]\right| \\
&\leq -\frac{1}{\lambda}\log\left[\left(\frac{\exp\left[-\lambda(M_y+uM_c)/(2n\kappa)\right]}{\exp\left[\lambda(M_y+uM_c)/(2n\kappa)\right]}\right)\left(\frac{\int_\Theta \exp\left[-\lambda\left(R_{n-i}(\theta)+uK_{n-i}(\theta)\right)\right]d\pi(\theta)}{\int_\Theta \exp\left[-\lambda\left(R_{n-i}'(\theta)+uK_{n-i}'(\theta)\right)\right]d\pi(\theta)}\right)\right] \\
&= \frac{M_y+uM_c}{n\kappa},
\end{aligned}$$

Thus, K_u satisfies the bounded differences property in Section 6.1 of Boucheron et al. (2013) with (in their notation) $c_i = (M_y+uM_c)/(n\kappa)$. By McDiarmid's inequality, (see McDiarmid (1989)), it holds that for any $t \geq 0$,

$$P^n\left(K_u(Z_1,\dots,Z_n) - E_{P^n}\left[K_u(Z_1,\dots,Z_n)\right] > t\right) \leq \exp\left(-\frac{2nt^2\kappa^2}{(M_y+uM_c)^2}\right).$$

Substituting $t = \sqrt{(M_y + uM_c)^2 \log(1/\varepsilon)/(2n\kappa^2)}$, we obtain that for for any $\varepsilon \in (0,1]$,

$$P^n\left(K_u(Z_1,\ldots,Z_n) > E_{P^n}\left[K_u(Z_1,\ldots,Z_n)\right] + \sqrt{\frac{(M_y+uM_c)^2\log(1/\varepsilon)}{2n\kappa^2}}\right) \leq \varepsilon.$$

Therefore (1.43) holds.

Next will show that

$$E_{P^n}\left[K_u(Z_1,\ldots,Z_n)\right] \leq \int_\Theta R(\theta)d\rho^*_{\lambda,u}(\theta) + u\int_\Theta K(\theta)d\rho^*_{\lambda,u}(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho^*_{\lambda,u},\pi). \qquad (1.44)$$

To do so, we follow arguments in Section 7 of Freund et al. (2004) with adjustments to suit our setting.

First note that by Corollary 1.A.1 (a),

$$\begin{aligned}\int_\Theta R(\theta)d\rho^*_{\lambda,u}(\theta) + &u\int_\Theta K(\theta)d\rho^*_{\lambda,u}(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho^*_{\lambda,u},\pi)\\ &= -\frac{1}{\lambda}\log\left[\int_\Theta \exp\left[-\lambda\left(R(\theta)+uK(\theta)\right)\right]d\pi(\theta)\right]. \qquad (1.45)\end{aligned}$$

Next, by Assumption 1.3.1 and the definitions of $R(\theta)$ and $K(\theta)$, it follows that

$$-M_y - uM_c \leq R(\theta) + uK(\theta) \leq M_y + uM_c,$$

for all $\theta \in \Theta$. For any $\delta > 0$, let

$$\mathscr{B}_i = \left\{\theta \in \Theta : -(M_y+uM_c) + i\delta \leq R(\theta) + uK(\theta) < -(M_y+uM_c) + (i+1)\delta\right\},$$

Then $\mathscr{B}_0,\ldots,\mathscr{B}_k$ with $k = \lfloor 2(M_y+uM_c)/\delta \rfloor$, form a partition of Θ. For $i \in \{0,\ldots,k\}$ such that $\pi(\mathscr{B}_i) > 0$, define

$$\tilde{\varepsilon}_i \equiv \frac{\int_{\mathscr{B}_i} R_n(\theta) + uK_n(\theta)d\pi(\theta)}{\pi(\mathscr{B}_i)},$$

Then, as π is independent of the sample by Assumption 1.3.3 and $E_{P^n}[R_n(\theta) + uK_n(\theta)] = R(\theta) + uK(\theta)$,

$$E_{P^n}[\tilde{\varepsilon}_i] = \frac{\int_{\mathscr{B}_i} R(\theta) + uK(\theta) d\pi(\theta)}{\pi(\mathscr{B}_i)} \leq -(M_y + uM_c) + (i+1)\delta.$$

Combining this with the fact that $R(\theta) + uK(\theta) > -(M_y + uM_c) + i\delta$ for $\theta \in \mathscr{B}_i$,

$$\begin{aligned}
\int_\Theta \exp[-\lambda(R(\theta) + uK(\theta))] d\pi(\theta) &\leq \sum \pi(\mathscr{B}_i) \exp[-\lambda(-(M_y + uM_c) + i\delta)] \\
&\leq \sum \pi(\mathscr{B}_i) \exp[-\lambda(E_{P^n}[\tilde{\varepsilon}_i - \delta])] \\
&= \exp[\lambda\delta] \sum \pi(\mathscr{B}_i) \exp[-\lambda(E_{P^n}[\tilde{\varepsilon}_i])],
\end{aligned}$$

where the sums above are to be understood as summing over all $i \in \{0, \ldots, k\}$ such that $\pi(\mathscr{B}_i) > 0$. Taking the logarithm of each side of this inequality and multiplying by $-1/\lambda$, we have

$$\begin{aligned}
&-\frac{1}{\lambda} \log\left[\int_\Theta \exp[-\lambda(R(\theta) + uK(\theta))] d\pi(\theta)\right] \\
&\geq -\delta - \frac{1}{\lambda} \log\left[\sum \pi(\mathscr{B}_i) \exp[-\lambda(E_{P^n}[\tilde{\varepsilon}_i])]\right] \\
&\geq -\delta - \frac{1}{\lambda} E_{P^n}\left[\log\left(\sum \pi(\mathscr{B}_i) \exp[-\lambda\tilde{\varepsilon}_i]\right)\right] && (1.46) \\
&= -\delta - \frac{1}{\lambda} E_{P^n}\left[\log\left(\sum \pi(\mathscr{B}_i) \exp\left[-\lambda \frac{\int_{\mathscr{B}_i} R_n(\theta) + uK_n(\theta) d\pi(\theta)}{\pi(\mathscr{B}_i)}\right]\right)\right] \\
&\geq -\delta - \frac{1}{\lambda} E_{P^n}\left[\log\left(\sum \pi(\mathscr{B}_i) \frac{\int_{\mathscr{B}_i} \exp[-\lambda(R_n(\theta) + uK_n(\theta))] d\pi(\theta)}{\pi(\mathscr{B}_i)}\right)\right] && (1.47) \\
&= -\delta - \frac{1}{\lambda} E_{P^n}\left[\log\left(\int_\Theta \exp[-\lambda(R_n(\theta) + uK_n(\theta))] d\pi(\theta)\right)\right] \\
&= -\delta + E_{P^n}[K_u(Z_1, \ldots, Z_n)] && (1.48)
\end{aligned}$$

In the above, (1.46) follows from an application of Jensen's inequality applied to the concave function

$$x \mapsto -\log\left(\sum_i \pi(\mathscr{B}_i) \exp[x_i]\right),$$

where the concavity of this function follows from Lemma 1.A.6. (1.47) follows from another application of Jensen's inequality now applied to the convex function $\exp(x)$. (1.48) follows from (1.42). δ was arbitrary, so this produces

$$-\frac{1}{\lambda}\log\left[\int_{\Theta}\exp\left[-\lambda\left(R(\theta)+uK(\theta)\right)\right]d\pi(\theta)\right]\geq E_{P^n}\left[K_u(Z_1,\ldots,Z_n)\right],$$

which, in light of (1.45), shows that (1.44) holds. (1.43) and (1.44) together yield that

$$P^n\Bigg(K_u(Z_1,\ldots,Z_n)$$
$$>\int_{\Theta}R(\theta)d\rho^*_{\lambda,u}(\theta)+u\int_{\Theta}K(\theta)d\rho^*_{\lambda,u}(\theta)+\frac{1}{\lambda}D_{\mathrm{KL}}(\rho^*_{\lambda,u},\pi)+\sqrt{\frac{(M_y+uM_c)^2\log(1/\varepsilon)}{2n\kappa^2}}\Bigg)\leq\varepsilon,$$

which produces the statement of the lemma upon taking the compliment. ■

1.A.2 Proofs for Section 1.3

Proofs for Subsection 1.3.1: Statistical Setting and Policy Maker's Problem

We will utilize the following lemma in the proof of Theorem 1.3.1.

Lemma 1.A.8 *Under the assumptions and setting of Theorem 1.3.1, let* $\delta_c^+(x)=\max(\delta_c(x),0)$ *and* $\delta_c^-(x)=\max(-\delta_c(x),0)$ *denote the positive and negative parts of* $\delta_c(x)$*, respectively. Define*

$$\beta(b)=E_Q[\delta_c(X)1\{\delta_y(X)>b\delta_c(X)\}],\ b\in\mathbb{R},$$

which is the expected budget of the non-stochastic treatment assigmnet rule $1\{\delta_y(x)>b\delta_c(x)\}$*.*
(i) Let $\eta_B=\inf\{b\geq 0:\beta(b)\leq B\}$ *.* $\beta(b)$ *is non-increasing in b and* $0\leq\eta_B<\infty$*.*

(ii) Let

$$a_1=\begin{cases}\frac{B-\beta(\eta_B)}{E_Q\left[\delta_c^+(X)1\{\delta_y(X)=\eta_B\delta_c(X)\}\right]} & \text{if } \beta\left(\eta_B\right)<B \text{ and } \eta_B>0,\\ 0 & \text{else,}\end{cases}$$

and

$$a_2 = \begin{cases} \frac{\beta(\eta_B)-B}{E_Q\left[\delta_c^-(X)1\{\delta_y(X)=\eta_B\delta_c(X)\}\right]} & \text{if } \beta(\eta_B) > B, \\ 0 & \text{else.} \end{cases}.$$

Then these are well defined probabilities in that $\beta(\eta_B) < B$ *and* $\eta_B > 0$ *implies*

$$E_Q[\delta_c^+(X)1\{\delta_y(X) = \eta_B\delta_c(X)\}] > 0,$$

$\beta(\eta_B) > B$ *implies*

$$E_Q[\delta_c^-(X)1\{\delta_y(X) = \eta_B\delta_c(X)\}] > 0,$$

and $a_1, a_2 \in [0,1]$. *Furthermore, for* f^* *defined as in Theorem 1.3.1 with* η_B, a_1, *and* a_2 *as above, when* $\beta(0) > B$ *it holds that*

$$E_Q[\delta_c(X)f_B^*(x)] = B.$$

Proof of Lemma 1.A.8.

Proof of (i): To show $\beta(b)$ is non-increasing in b, write

$$\beta(b) = E_Q\left[\delta_c^+(X)1\{\delta_y(X) > b\delta_c(X)\}\right] - E_Q\left[\delta_c^-(X)1\{\delta_y(X) > b\delta_c(X)\}\right], \tag{1.49}$$

By definition of $\delta_c^+(x)$ and $\delta_c^-(x)$,

$$\delta_c^+(x)1\{\delta_y(x) - b\delta_c(x)\}$$

is non-increasing in b and

$$\delta_c^-(x)1\{\delta_y(x) - b\delta_c(x)\}$$

is non-decreasing in b for all $x \in \mathscr{X}$. It follows that $\beta(b)$ is non-increasing in b.

Checking $0 \leq \eta_B < \infty$ translates to verifying that our form of policy assignment rule can meet the budget requirement. Let $\{b_n\}$ be any non-negative sequence such that $b_n \to \infty$. Then, $E_Q|\delta_c(X)| < \infty$ and $E_Q|\delta_y(X)| < \infty$, equation (1.49), and an application of the dominated convergence theorem yield

$$\begin{aligned}\lim_{n\to\infty} \beta(b_n) &= \lim_{n\to\infty} E_Q\left[\delta_c^+(X)1\{\delta_y(X) > b_n\delta_c(X)\}\right] - \lim_{n\to\infty} E_Q\left[\delta_c^-(X)1\{\delta_y(X) > b_n\delta_c(X)\}\right]\\ &= 0 - E_Q[\delta_c^-(X)] < B.\end{aligned}$$

The inequality follows from the assumption that $B > E_Q[\delta_c(X)1\{\delta_c(X) < 0\}] = -E_Q[\delta_c^{-1}(X)]$. As $\beta(b)$ is non-increasing, we have either $\{b \geq 0 : \beta(b) \leq B\} = [r, \infty)$ or $\{b \geq 0 : \beta(b) \leq B\} = (r, \infty)$ for some $r \in \mathbb{R}_{\geq 0} = \{x \in \mathbb{R} : x \geq 0\}$. It follows that $0 \leq \eta_B < \infty$.

Proof of (ii): Let $b \in \mathbb{R}$. For any sequence $b_n \uparrow b$, by the dominated convergence theorem we have

$$\begin{aligned}\lim_{b_n\uparrow b} \beta(b_n) =& \lim_{b_n\uparrow b} E_Q[\delta_c^+(X)1\{\delta_y(X) > b_n\delta_c(X)\}] - \lim_{b_n\uparrow b} E_Q[\delta_c^-(X)1\{\delta_y(X) > b_n\delta_c(X)\}]\\ =& E_Q[\delta_c^+(X)1\{\delta_y(X) \geq b\delta_c(X)\}] - E_Q[\delta_c^-(X)1\{\delta_y(X) > b\delta_c(X)\}].\\ =& E_Q[\delta_c^+(X)1\{\delta_y(X) > b\delta_c(X)\}] + E_Q[\delta_c^+(X)1\{\delta_y(X) = b\delta_c(X)\}]\\ & - E_Q[\delta_c^-(X)1\{\delta_y(X) > b\delta_c(X)\}].\\ =& \beta(b) + E_Q[\delta_c^+(X)1\{\delta_y(X) = b\delta_c(X)\}]\end{aligned}$$

This yields

$$\lim_{x\to b^-} \beta(x) = \beta(b) + E_Q\left[\delta_c^+(X)1\left\{\delta_y(X) = b\delta_c(X)\right\}\right]. \tag{1.50}$$

Similar steps now starting with any sequence $b_n \downarrow b$ produce that

$$\lim_{x\to b^+} \beta(x) = \beta(b) - E_Q\left[\delta_c^-(X)1\left\{\delta_y(X) = b\delta_c(X)\right\}\right]. \tag{1.51}$$

As $\beta(\cdot)$ is non-increasing, it has at most countably many discontinuities, which occur at values

b for which either $E_Q[\delta_c^+(X)1\{\delta_y(X) = b\delta_c(X)\}] > 0$ or $E_Q[\delta_c^-(X)1\{\delta_y(X) = b\delta_c(X)\}] > 0$ or both.

Now, if $B > \beta(\eta_B)$ and $\eta_B > 0$, by definition of η_B we have that $\beta(\eta') > B$ for any $\eta' < \eta_B$. Combined with (1.50), we obtain

$$\beta\left(\eta_B\right) < B \leq \beta\left(\eta_B\right) + E_Q\left[\delta_c^+ 1\{\delta_y(X) = \eta_B\delta_c(X)\}\right],$$

which implies that $E_Q[\delta_c^+(X)1\{\delta_y(X) = \eta_B\delta_c(X)\}] > 0$ and $a_1 \in [0,1]$.

Next, if $B < \beta(\eta_B)$, by defnition of η_B we have $\beta(\eta') \leq B$ for any $\eta' > \eta_B$. Combining this with (1.51), we obtain

$$\beta(\eta_B) - E_Q\left[\delta_c^-(X)\,1\left\{\delta_y(X) = \eta_B\delta_c(X)\right\}\right] \leq B < \beta\left(\eta_B\right).$$

This implies $E_Q[\delta_c^-(X)1\{\delta_y(X) = \eta_B\delta_c(X)\}] > 0$ and $a_2 \in [0,1]$.

For the last claim of (ii), write

$$\begin{aligned} E_Q\left[\delta_c(X)f_B^*(x)\right] =& E_Q\left[\delta_c(X)1\left\{\delta_y(X) > \eta_B\delta_c(X)\right\}\right] \\ &+ a_1E_Q\left[\delta_c(X)1\left\{\delta_y(X) = \eta_B\delta_c(X)\right\}1\left\{\delta_c(X) > 0\right\}\right] \\ &+ a_2E_Q\left[\delta_c(X)1\left\{\delta_y(X) = \eta_B\delta_c(X)\right\}1\left\{\delta_c(X) < 0\right\}\right]. \end{aligned} \tag{1.52}$$

When $\beta(0) > B$, there are 3 scenarios for $\beta(\eta_B)$: (i) $\beta(\eta_B) = B$ and $\eta_B > 0$; (ii) $\beta(\eta_B) < B$ and $\eta_B > 0$; or (iii) $\beta(\eta_B) > B$ and $\eta_B \geq 0$. For scenario (i), we have $a_1 = a_2 = 0$ and the result holds as $E_Q[\delta_c(X)1\{\delta_y(X) > \eta_B\delta_c(X)\}] = \beta(\eta_B) = B$. For scenario (ii), $a_2 = 0$ and (1.52) becomes

$$\begin{aligned} E_Q\left[\delta_c(X)f_B^*(x)\right] =& E_Q\left[\delta_c(X)1\left\{\delta_y(X) > \eta_B\delta_c(X)\right\}\right] \\ &+ \frac{B - \beta(\eta_B)}{E_Q\left[\delta_c^+(X)1\left\{\delta_y(X) = \eta_B\delta_c(X)\right\}\right]}E_Q\left[\delta_c^+(X)1\left\{\delta_y(X) = \eta_B\delta_c(X)\right\}\right] \\ =& B. \end{aligned}$$

For scenario (iii), $a_1 = 0$ and (1.52) becomes

$$\begin{aligned}E_Q\left[\delta_c(X)f_B^*(x)\right] =& E_Q\left[\delta_c(X)1\left\{\delta_y(X) > \eta_B\delta_c(X)\right\}\right]\\&-\frac{\beta(\eta_B)-B}{E_Q\left[\delta_c^-(X)1\left\{\delta_y(X) = \eta_B\delta_c(X)\right\}\right]}E_Q\left[\delta_c^-(X)1\left\{\delta_y(X) = \eta_B\delta_c(X)\right\}\right]\\&= B.\end{aligned}$$

This completes the proof of (ii). ■

Proof of Theorem 1.3.1.

The existence of $\eta_B \geq 0$, $a_1, a_2 \in [0,1]$ such that either $\eta_B = a_1 = a_2 = 0$ (then f^* simplifies to f^*) when $K(f^*) \leq B$ or else (η_B, a_1, a_2) are such that $K(f^*) = B$ when $K(f^*) > B$ follows from Lemma 1.A.8. To see this note $\beta(0) = K(f^*)$, where $\beta(\cdot)$ is defined in Lemma 1.A.8. Thus, the statement about the budget being used entirely when $K(f^*) > B$ is stated directly in Lemma 1.A.8. When $\beta(0) = K(f^*) \leq B$, η_B as defined in Lemma 1.A.8 is equal to zero and then both $a_1 = a_2 = 0$ also from their definitions there.

Next we need to verify that f^* satisfies (1.6), i.e. is an optimal budget-constrained treatment policy. Let $r : \mathscr{X} \to [0,1]$ denote any other stochastic treatment assignment rule that satisfies the budget constraint $K(r) \leq B$. As in Sun et al. (2021), we proceed by verifying that

$$E_Q\left[\delta_y(X)f_B^*(x)\right] \geq E_Q\left[\delta_y(X)r(X)\right].$$

By the definition of f^*, when $\delta_y(x) > \eta_B\delta_c(x)$ we also have $f_B^*(x) - r(x) \geq 0$. Hence $\delta_y(x)(f_B^*(x) - r(x)) \geq \eta_B\delta(x)(f_B^*(x) - r(x))$ in this case. When $\delta_y(x) < \eta_B\delta_c(x)$, we have $f_B^*(x) - r(x) \leq 0$ and hence $\delta_y(x)(f_B^*(x) - r(x)) \geq \eta_B\delta(x)(f_B^*(x) - r(x))$ in this case as well. It follows that

$$E_Q\left[\delta_y(X)\left(f_B^*(x) - r(X)\right)\right] \geq \eta_B E_Q\left[\delta_c(X)\left(f_B^*(x) - r(X)\right)\right]. \tag{1.53}$$

There are two possible scenarios: $K(f^*) \leq B$ or else $K(f^*) > B$. When $K(f^*) \leq B$, we have $\eta_B = 0$ and hence the right-hand-side of (1.53) is zero implying f^* is optimal. If, alternatively, $K(f^*) > B$, then we know that $K(f^*) = B$ and $K(r) \leq B$. Thus

$$E_Q[\delta_c(X)(f_B^*(x) - r(X))] = K(f^*) - K(r) \geq 0.$$

Now the right-hand-side of (1.53) is non-negative (as $\eta_B \geq 0$) and f^* is again optimal.

Lastly we need to show that if

$$E_Q[1\{\delta_y(X) = \eta_B \delta_c(X)\}] = 0, \tag{1.54}$$

then f^* is deterministic and unique (in an almost sure sense). It is clear from the form of f^* that it is almost surely equivalent to $1\{\delta_y(x) > \eta_B \delta_c(x)\}$ in this setting. Additionally, with the choices of η_B, a_1, a_2 given in Lemma 1.A.8 this will be true for all $x \in \mathscr{X}$. To see that this follows from the proof of Lemma 1.A.8, by (1.50) and (1.51) there, $\beta(b)$ is continuous in this scenario so that $\beta(\eta_B) = B$ when $\eta_B > 0$; this implies $a_1 = a_2 = 0$ when $\eta_B > 0$. When $\eta_B = 0$, we must have $\beta(0) \leq B$ and then again $a_1 = a_2 = 0$ as defined in Lemma 1.A.8.

To check uniqueness, let $r(x)$ be any other treatment assignment rule that satisfies the budget constraint $K(r) \leq B$ and is not a.s. equal to $f_B^*(x)$. When $\eta_B = 0$, r must then assign treatment for a subset of $\mathscr{X}$ with positive probability that has negative CATE or else fail to assign treatment to some subset of $\mathscr{X}$ that has positive CATE with positive probability (or both). This results in lower expected welfare than f^*, so r cannot be optimal. When $\eta_B > 0$, the argument is similar to that showing f^* is optimal. When $\eta_B > 0$, its definition in Lemma 1.A.8 indicates that $K(f^*) = \beta(0) > B$ (and from (1.52) in Lemma 1.A.8, it follows that $\eta_B > 0$ in this case must be the unique choice for which the expected budget of f^* is B). $P(f_B^*(x) \neq r(x)) > 0$ then implies that for some subset of $\mathscr{X}$ with positive probability we must have $f_B^*(x) - r(x) > 0$ when $\delta_y(x) > \eta_B \delta_c(x)$ or else $f_B^*(x) - r(x) < 0$ when $\delta_y(x) < \eta_B \delta_c(x)$ (or both). This implies that the inequality in (1.53) is strict. As f^* uses up the entire budget (1.53) now implies the left-hand-side is strictly positive, which concludes the proof. ■

Proofs for Subsection 1.3.3: Initial Properties of the Gibbs Posterior

Proof of Lemma 1.3.1. First we derive the result in (1.17). There are two possible scenarios. First, if $\Lambda(0) \leq B$, i.e. the "cost" at $u = 0$ is within budget, then $\tilde{\rho}_{A,H,\lambda,0} \in \mathscr{E}_B$ and $\tilde{\rho}_{A,H,\lambda,0} = \rho_{\lambda A,\pi}$ in the notation of Corollary 1.A.1. Then the result follows from Corollary 1.A.1 (a). Note that this scenario captures the case when $B = \infty$, i.e. when there is no budget constraint.

In the second scenario, $\Lambda(0) > B$ (and $B < \infty$). Assume this is case for the remainder of the proof of property (1.17). First, we will show that this implies $\Lambda(u)$ is (strictly) decreasing in u and that there exists a unique $\overline{u}_B > 0$ such that $\Lambda(\overline{u}_B) = B$. Note below that at any point $u \geq 0$, because the derivatives of the integrands are dominated by integrable functions on intervals of the form $(u-a, u+b)$, some $a, b > 0$, and as $\Lambda(u)$ is easily extended in definition to negative values of u in neighborhoods of 0, we can exchange differentiation and integration. We have

$$\begin{aligned}
&\frac{d}{du}\Lambda(u) \\
&= \frac{d}{du}\left[\left(\int_\Theta H(\theta)\exp[-\lambda(A(\theta)+uH(\theta))]\,d\pi(\theta)\right)\left(\int_\Theta \exp[-\lambda(A(\theta)+uH(\theta))]\,d\pi(\theta)\right)^{-1}\right] \\
&= -\lambda\int_\Theta H^2(\theta)\,d\tilde{\rho}_{A,H,\lambda,u}(\theta) + \lambda\left(\int_\Theta H(\theta)\,d\tilde{\rho}_{A,H,\lambda,u}(\theta)\right)^2 \\
&= -\lambda\mathbb{V}_{\theta\sim\tilde{\rho}_{A,H,\lambda,u}}[H(\theta)] \\
&< 0, \qquad (1.55)
\end{aligned}$$

where $\mathbb{V}_{\theta\sim\tilde{\rho}_{A,H,\lambda,u}}[H(\theta)]$ denotes the variance of $H(\theta)$ when $\theta \sim \tilde{\rho}_{A,H,\lambda,u}$. Note the strict inequality of the last line holds because the distribution of $H(\theta)$ induced by $\tilde{\rho}_{A,H,\lambda,u}$ is degenerate only when the distribution of $H(\theta)$ induced by π is degenerate. If this were the case, (1.16) would imply that $\Lambda(0) < B$. Hence the strict inequality when $\Lambda(0) \geq B$, which includes our current $\Lambda(0) > B$ scenario.

Now, note that (1.16) implies there exist $\varepsilon_1, \eta > 0$ such that $\pi(\{\theta : H(\theta) \leq B - \varepsilon_1\}) =$

$\eta > 0$. Let ε_2 be such that $0 < \varepsilon_2 < \varepsilon_1$. Letting M_h and M_a be such that $|H(\theta)| \leq M_h$ and $|A(\theta)| \leq M_a$ for all θ (as these functions are assumed bounded), we have

$$\begin{aligned}
&\int_\Theta H(\theta)\, d\tilde{\rho}_{A,H,\lambda,u}(\theta) \\
&\leq (B-\varepsilon_2) + M_h \int_\Theta 1\{H(\theta) > B-\varepsilon_2\}\, d\tilde{\rho}_{A,H,\lambda,u}(\theta) \\
&= (B-\varepsilon_2) + M_h \frac{\int_\Theta 1\{H(\theta) > B-\varepsilon_2\} \exp[-\lambda(A(\theta)+uH(\theta))]\, d\pi(\theta)}{\int_\Theta (1\{H(\theta) \leq B-\varepsilon_1\} + 1\{H(\theta) > B-\varepsilon_1\}) \exp[-\lambda(A(\theta)+uH(\theta))]\, d\pi(\theta)} \\
&\leq (B-\varepsilon_2) + M_h \frac{\int_\Theta 1\{H(\theta) > B-\varepsilon_2\} \exp[-\lambda(A(\theta)+uH(\theta))]\, d\pi(\theta)}{\int_\Theta 1\{H(\theta) \leq B-\varepsilon_1\} \exp[-\lambda(A(\theta)+uH(\theta))]\, d\pi(\theta)} \\
&\leq (B-\varepsilon_2) + M_h \left(\frac{\exp[-\lambda u(B-\varepsilon_2)]}{\exp[-\lambda u(B-\varepsilon_1)]}\right)\left(\frac{\exp[\lambda M_a]}{\eta \exp[-\lambda M_a]}\right) \\
&= (B-\varepsilon_2) + \exp[-\lambda u(\varepsilon_1-\varepsilon_2)]\left(\frac{M_h \exp[2\lambda M_a]}{\eta}\right).
\end{aligned}$$

As $\varepsilon_1 - \varepsilon_2 > 0$, for large enough values of u it holds that $\Lambda(u) < B$. Then, as $\Lambda(u)$ is continuous and strictly decreasing in u it follows that there is a unique $\overline{u}_B > 0$ such that $\Lambda(\overline{u}_B) = B$.

To finish the proof the property in (1.17), we need to show that when $\Lambda(0) > B$, $\tilde{\rho}_{A,H,\lambda,\overline{u}_B}$ is the optimal probability measure on Θ for the minimization problem. Replacing A in Corollary 1.A.1 (a) with the $A + \overline{u}_B H$ as given above and noting that $\tilde{\rho}_{A,H,\lambda,\overline{u}_B} = \rho_{\lambda(A+\overline{u}_B H),\pi}$, we have that for any $\rho \in \mathscr{E}_B$,

$$\begin{aligned}
&\tilde{\rho}_{A,H,\lambda,\overline{u}_B} \\
&= \operatorname*{arg\,min}_{\rho \in \mathscr{P}_\pi(\Theta)} \left[\int_\Theta \{A(\theta) + \overline{u}_B H(\theta)\}\, d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi)\right] && (1.56) \\
&= \operatorname*{arg\,min}_{\rho \in \mathscr{P}_\pi(\Theta)} \left[\int_\Theta \{A(\theta) + \overline{u}_B H(\theta)\}\, d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) - \overline{u}_B B\right] \\
&= \operatorname*{arg\,min}_{\mathscr{E}_{H,B}} \left[\int_\Theta A(\theta)\, d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) + \overline{u}_B \left(\int_\Theta H(\theta)\, d\rho(\theta) - B\right)\right] && (1.57) \\
&= \operatorname*{arg\,min}_{\{\rho \in \mathscr{P}_\pi(\Theta) : \int_\Theta H(\theta) d\rho(\theta) = B\}} \left[\int_\Theta A(\theta)\, d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) + \overline{u}_B \left(\int_\Theta H(\theta)\, d\rho(\theta) - B\right)\right], && (1.58)
\end{aligned}$$

where the third equality holds in our specific setting because $\tilde{\rho}_{A,H,\lambda,\overline{u}_B} \in \mathscr{E}_{H,B}$ and $\mathscr{E}_{H,B} \subset \mathscr{P}_\pi(\Theta)$. The fourth equality follows similar reasoning. Next note that for any $\rho \in \mathscr{E}_{H,B}$, as $\overline{u}_B > 0$,

$$\int_\Theta A(\theta)\, d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) \geq \int_\Theta A(\theta)\, d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) + \overline{u}_B \left(\int_\Theta H(\theta)\, d\rho(\theta) - B \right) \tag{1.59}$$

$$\geq \int_\Theta A(\theta)\, d\tilde{\rho}_{A,H,\lambda,\overline{u}_B}(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}\left(\tilde{\rho}_{A,H,\lambda,\overline{u}_B}, \pi\right), \tag{1.60}$$

where (1.60) follows from (1.57) and the fact that $\int_\Theta H(\theta) d\tilde{\rho}_{A,H,\lambda,\overline{u}_B}(\theta) = B$. Because the inequality in (1.59) is strict whenever $\int_\Theta H(\theta) d\rho < B$ it follows from (1.58) that $\tilde{\rho}_{A,H,\lambda,\overline{u}_B}$ is the argmin when $\Lambda(0) > B$, completing the proof of the property in (1.17).

Next we need to prove the property in (1.18). This property is trivial when $B = \infty$. Assume $B < \infty$ for the remainder of the proof. Let

$$h(u) = \int_\Theta A(\theta) d\tilde{\rho}_{A,H,\lambda,u}(\theta) + u \left(\int_\Theta H(\theta) d\tilde{\rho}_{A,H,\lambda,u}(\theta) - B \right) + \frac{1}{\lambda} D_{\mathrm{KL}}\left(\tilde{\rho}_{A,H,\lambda,u}, \pi\right).$$

By the definition of $\overline{u}_B$, we need to show that the supremum of $h(u)$ over $u \geq 0$ is achieved at $\overline{u}_B$. Observe that by Corollary 1.A.1 (a), as $\tilde{\rho}_{A,H,\lambda,u} = \rho_{\lambda(A+uH),\pi}$ in the notation there,

$$\begin{aligned} &\int_\Theta \{A(\theta) + uH(\theta)\}\, d\tilde{\rho}_{A,H,\lambda,u}(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\tilde{\rho}_{A,H,\lambda,u}, \pi) \\ &= -\frac{1}{\lambda} \log \left[\int_\Theta \exp\left[-\lambda\left(A(\theta) + uH(\theta)\right)\right] d\pi(\theta) \right]. \end{aligned} \tag{1.61}$$

Utilizing this it is straightforward to derive that

$$\begin{aligned} \frac{d}{du} h(u) &= \frac{\int_\Theta H(\theta) \exp\left[-\lambda\left(A(\theta) + uH(\theta)\right)\right] d\pi(\theta)}{\int_\Theta \exp\left[-\lambda\left(A(\theta') + uH(\theta')\right)\right] d\pi(\theta')} - B \\ &= \Lambda(u) - B, \end{aligned} \tag{1.62}$$

where we may exchange differentiation and expectation following similar reasoning as before.

In the proof of the property in (1.17) it is shown that $\Lambda(u)$ is strictly decreasing on $[0,\infty)$ when $\Lambda(0) \geq B$. When $\Lambda(0) > B$, by the definition of $\bar{u}_B$ we have $\Lambda(\bar{u}_B) = B$ with $\bar{u}_B > 0$. It follows that the supremum of $h(u)$, which is continuous in u, is achieved at $\bar{u}_B$. This is because, from (1.62), the derivative of $h(u)$ is positive on $[0,\bar{u}_B)$, zero at $\bar{u}_B$ and decreasing on $(\bar{u}_B,\infty)$. If $\Lambda(0) = B$, we have that the supremum is achieved at 0, which is $\bar{u}_B$ in this case by the definition $\bar{u}_B$, as the derivative of $h(u)$ is now zero at $u = \bar{u}_B = 0$ and negative for $u \in (0,\infty)$. Conversely, if $\Lambda(0) < B$, nearly identical steps to those in the proof of the property in (1.17) show that $\Lambda(u)$ is non-increasing in u for $u \geq 0$. Hence in this case the derivative of $h(u)$ is negative for $u \in [0,\infty)$ and the supremum is achieved at 0, which by definition, is the value of $\bar{u}_B$ when $\Lambda(0) \leq B$. ■

Proof of Lemma 1.3.2. Part (a). Given Assumptions 1.3.2 and 1.3.4 (i) for $B \in \mathbb{R} \cup \{\infty\}$, this is an immediate corollary of Lemma 1.3.1 taking $A(\theta) = R_n(\theta)$ and $H(\theta) = K_n(\theta)$.

Part (b). Again let $A(\theta) = R_n(\theta)$, $H(\theta) = K_n(\theta)$, and $\tilde{\rho}_{A,H,\lambda,u} = \hat{\rho}_{\lambda,u}$ in the notation of Lemma 1.3.1. Observe that, for a fixed sample S, as the distribution of $K_n(\theta)$ induced by $\theta \sim \hat{\rho}_{\lambda,u}$ is degenerate only when the distribution of $K_n(\theta)$ induced by π is degenerate, which is assumed to not be the case (with probability one) by Assumption 1.3.4 (ii), P^n almost surely it holds that

$$\int_{\Theta} K_n(\theta) d\hat{\rho}_{\lambda,u}$$

cannot take any value b that does not satisfy

$$\pi\left(\{\theta : K_n(\theta) < b\}\right) > 0.$$

It follows that, P^n almost surely,

$$\pi\left(\left\{\theta : K_n(\theta) < \widehat{B}(\hat{\rho}_{\lambda,u})\right\}\right) > 0.$$

Then, the result for part (b) follows by applying Lemma 1.3.1 with $A(\theta) = R_n(\theta), H(\theta) = K_n(\theta)$, and $B = \widehat{B}(\hat{\rho}_{\lambda,u})$. ■

1.A.3 Proofs for Section 1.4

The proofs of Theorems 1.4.2 and 1.4.3 will utilize the following lemmas that follow from Lemma 1.3.1. We again utilize the notation in (1.19) and (1.20) for $\mathscr{E}_B$ and $\widehat{\mathscr{E}}_B$, respectively.

Lemma 1.A.9 *(a) Let Assumptions 1.3.2 and 1.3.4 (i) hold for $B \in \mathbb{R}$. For any $\lambda > 0$ and $B' \geq B$, P^n almost surely it holds that*

$$\begin{aligned}&\min_{\widehat{\mathscr{E}}_{B'}} \left[\int_\Theta R_n(\theta) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) \right] \\ &= \sup_{u \geq 0} \left[\int_\Theta R_n(\theta) d\hat{\rho}_{\lambda,u}(\theta) + u \left(\int_\Theta K_n(\theta) d\hat{\rho}_{\lambda,u}(\theta) - B' \right) + \frac{1}{\lambda} D_{\mathrm{KL}} \left(\hat{\rho}_{\lambda,u}, \pi \right) \right].\end{aligned}$$

(b) Let Assumptions 1.3.2 and 1.3.4 (i) hold for $B \in \mathbb{R}$. The following properties hold P^n almost surely. For any $\lambda > 0$ and $B' \geq B$, $u^*(B', \lambda)$ exist, is unique, and satisfies that $u^*(B', \lambda) = 0$ when $\int_\Theta K(\theta) d\rho^*_{\lambda,0}(\theta) \leq B'$ whereas, when $\int_\Theta K(\theta) d\rho^*_{\lambda,0}(\theta) > B'$, $u^*(B', \lambda)$ is positive and satisfies $\int_\Theta K(\theta) d\rho^*_{\lambda,u^*(B',\lambda)}(\theta) = B'$. Additionally,

$$\rho^*_{\lambda,u^*(B',\lambda)} = \arg\min_{\mathscr{E}_{B'}} \left[\int_\Theta R(\theta) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) \right],$$

and

$$\begin{aligned}&\min_{\mathscr{E}_{B'}} \left[\int_\Theta R(\theta) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) \right] \\ &= \sup_{u \geq 0} \left[\int_\Theta R(\theta) d\rho^*_{\lambda,u}(\theta) + u \left(\int_\Theta K(\theta) d\rho^*_{\lambda,u}(\theta) - B' \right) + \frac{1}{\lambda} D_{\mathrm{KL}} \left(\rho^*_{\lambda,u}, \pi \right) \right].\end{aligned}$$

(c) Let Assumptions 1.3.2 and 1.3.4 (ii) hold. For any $B' \geq B(\hat{\rho}_{\lambda,u})$, the following event

occurs P^n almost surely

$$\begin{aligned}
&\min_{\rho \in \mathscr{E}_{B'}} \left[\int_\Theta R(\theta) d\rho(\theta) + \frac{1}{\lambda} D_{\text{KL}}(\rho, \pi) \right] \\
&= \sup_{a \geq 0} \left[\int_\Theta R(\theta) d\rho^*_{\lambda,a}(\theta) + a \left(\int_\Theta K(\theta) d\rho^*_{\lambda,a}(\theta) - B' \right) + \frac{1}{\lambda} D_{\text{KL}}\left(\rho^*_{\lambda,a}, \pi\right) \right].
\end{aligned}$$

Proof of Lemma 1.A.9. Part (a). When

$$\pi(\theta \in \Theta : K_n(\theta) < B) > 0$$

holds for B, it also holds for $B' \geq B$. Therefore part (a) follows from Lemma 1.3.1 with $A(\theta) = R_n(\theta)$, $H(\theta) = K_n(\theta)$, $\tilde{\rho}_{A,H,\lambda,u} = \hat{\rho}_{\lambda,u}$ and combining the statements in (1.17) and (1.18).

Part (b). When

$$\pi(\theta \in \Theta : K(\theta) < B) > 0$$

holds for B, it also holds for $B' \geq B$. Then the result follows from Lemma 1.3.1 with $A(\theta) = R(\theta)$, $H(\theta) = K(\theta)$, and $\tilde{\rho}_{A,H,\lambda,u} = \rho^*_{\lambda,u}$.

Part (c). Note that by Assumption 1.3.4 (ii), as $\hat{\rho}_{\lambda,u}$ and π are each absolutely continuous with respect to the other, we have that the distribution of $K(\theta)$ induced by $\theta \sim \hat{\rho}_{\lambda,u}$ is not degenerate. Therefore,

$$\pi\left(\left\{\theta : K(\theta) < \int_\Theta K(\theta) d\hat{\rho}_{\lambda,u} = B(\hat{\rho}_{\lambda,u})\right\}\right) > 0.$$

It follows that for any $B' \geq B(\hat{\rho}_{\lambda,u})$, we have

$$\pi\left(\{\theta : K(\theta) < B'\}\right) > 0.$$

Then, the result of part (c) follows from applying Lemma 1.3.1 with $A(\theta) = R(\theta)$, $H(\theta) = K(\theta)$, and $\tilde{\rho}_{A,H,\lambda,u} = \rho^*_{\lambda,u}$ and then combining equations (1.17) and (1.18) there. ■

Proofs for Subsection 1.4.1: Regret Bounds and Oracle-Type Inequalities

Proof of Theorem 1.4.1. Part (a). When $V_n(\theta) = R_n(\theta)$, $V(\theta) = R(\theta)$, and $M_\ell = M_y$, we have the setup for Theorem 1.A.1 with

$$\ell_v(Z,\theta) = \left(\frac{YD}{e(X)} - \frac{Y(1-D)}{1-e(X)}\right)(f^*(X) - f_\theta(X)),$$

$L(\theta) = V(\theta)$, and $L_n(\theta) = V_n(\theta)$. Note that, by Assumption 1.3.1, parts (iii) and (iv), we have that $-M_\ell/2\kappa \leq \ell_v(Z,\theta) \leq M_\ell/2\kappa$ a.s. Similarly, when $V_n(\theta) = K_n(\theta)$, $V(\theta) = K(\theta)$, and $M_\ell = M_c$, we have the setup for Theorem 1.A.1 now with

$$\ell_v(Z,\theta) = \left(\frac{CD}{e(X)} - \frac{C(1-D)}{1-e(X)}\right) f_\theta(X),$$

and again taking $L(\theta) = V(\theta)$ and $L_n(\theta) = V_n(\theta)$. Again Assumption 1.3.1, parts (iii) and (iv), yields that $-M_\ell/2\kappa \leq \ell_v(Z,\theta) \leq M_\ell/2\kappa$ a.s.

Given this setup, we apply Theorem 1.A.1 in the same way for either of the settings for $L(\theta), L_n(\theta)$ and M_ℓ. We need an appropriate choice for $D(\cdot,\cdot)$ and to then verify the condition in (1.31). Importantly, in either setting we have that, for any $\theta \in \Theta$, $\ell_v(Z_1,\theta),\ldots,\ell_v(Z_n,\theta)$ is an iid set of random variables taking values in $[-M_\ell/2\kappa, M_\ell/2\kappa]$ almost surely. For either $s \in \{-1,1\}$, take $D[L_n(\theta),L(\theta)] = s(L_n(\theta) - L(\theta))$. We need to verify the condition in (1.31) and determine an appropriate $f(\lambda,n)$. Start with $s = 1$. Then, by Hoeffding's lemma (see, for example, Massart

(2007), page 21), for any $\theta \in \Theta$,

$$\begin{aligned} E_{P^n}\left[\exp\left(\lambda\left[L_n(\theta) - L(\theta)\right]\right)\right] &= E_{P^n}\left[\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n}\left(\ell_v(Z_i,\theta) - E_P\left[\ell_v(Z_i,\theta)\right]\right)\right)\right] \\ &= \prod_{i=1}^{n} E_P\left[\exp\left\{\frac{\lambda}{n}\left(\ell_v(Z_i,\theta) - E_P\left[\ell_v(Z_i,\theta)\right]\right)\right\}\right] \\ &\le \prod_{i=1}^{n}\exp\left(\frac{\lambda^2 M_\ell^2}{8\kappa^2 n^2}\right) = \exp\left(\frac{\lambda^2 M_\ell^2}{8\kappa^2 n}\right) \end{aligned} \tag{1.63}$$

Nearly identical steps in the $s = -1$ case, now applying Hoeffding's lemma to $-\ell_v(Z_i,\theta)$ produce that

$$E_{P^n}\left[\exp\left(\lambda\left[L(\theta) - L_n(\theta)\right]\right)\right] \le \exp\left(\frac{\lambda^2 M_\ell^2}{8\kappa^2 n}\right). \tag{1.64}$$

Integrating with respect to π, (1.63) and (1.64) yield that

$$\int_\Theta E_{P^n}\left[\lambda s\left(R_n(\theta) - R(\theta)\right)\right] d\pi(\theta) \le \exp\left(\frac{\lambda^2 M_\ell^2}{8\kappa^2 n}\right),\ \ s \in \{-1, 1\}.$$

We can reverse the order of integration on the left-hand of the above inequality, as π is independent of the sample by Assumption 1.3.3. Therefore, condition (1.31) in Theorem 1.A.1 holds with $f(\lambda, n) = \lambda^2 M_\ell^2/(8n\kappa^2)$. Applying Theorem 1.A.1 completes the proof for Part (a).

Part (b). We utilize the same notation in terms of $\ell_v(Z,\theta)$ in the two scenarios for $V_n(\theta)$, $V(\theta)$, and M_ℓ as in part (a). Let E_1 denote the event that the following inequality holds,

$$D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,u}, \rho^*_{\lambda,u}\right) \le \frac{\lambda\sqrt{2}\left(M_y + uM_c\right)}{\kappa\sqrt{n}}\sqrt{\log\left(2\sqrt{n}\right) + \log\frac{2}{\varepsilon}} + \frac{\lambda^2\left(M_y + uM_c\right)^2}{2n\kappa^2}. \tag{1.65}$$

Note that by Lemma 1.A.3, $P^n(E_1) \ge 1 - \varepsilon/2$.

Next, let E_2 denote the event that the following inequality holds,

$$\left(\int_\Theta \left[V_n(\theta) - V(\theta)\right] d\hat{\rho}_{\lambda,u}(\theta)\right)^2 \le \frac{M_\ell^2}{2n\kappa^2}\left[D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,u}, \rho^*_{\lambda,u}\right) + \log\left(2\sqrt{n}\right) + \log\frac{2}{\varepsilon}\right]. \tag{1.66}$$

In the setup of Theorem 1.A.2 (b), take

$$\ell(Z,\theta) = \left(\ell_v(Z,\theta) + \frac{M_\ell}{2\kappa}\right)\left(\frac{\kappa}{M_\ell}\right).$$

Then, for any $\theta \in \Theta$, $\ell(Z,\theta) \in [0,1]$ (P almost surely). Applying Theorem 1.A.2 (b) yields that $P^n(E_2) \geq 1 - \varepsilon/2$.

Then, the following a union bound argument,

$$\begin{aligned} P^n\left(E_1 \cap E_2\right) &= 1 - P^n\left(E_1^c \cup E_2^c\right) \\ &\geq 1 - P^n\left(E_1\right) - P^n\left(E_2\right) \\ &\geq 1 - \frac{\varepsilon}{2} - \frac{\varepsilon}{2} = 1 - \varepsilon, \end{aligned}$$

yields that events E_1 and E_2 occur jointly with probability greater than $1-\varepsilon$. In the intersection of these events, plugging (1.65) into (1.66) produces the result in part (b).

Part (c). The proof follows similar steps to that in part (b). Define E_1 the same way as in part (b). Now, E_2 is defined to be the event that

$$\int_\Theta s\left[V_n(\theta) - V(\theta)\right] d\hat{\rho}_{\lambda,u}(\theta) \leq \frac{1}{\lambda} D_{\mathrm{KL}}(\rho,\pi) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_\ell^2}{8n\kappa^2} + \log\frac{2}{\varepsilon}\right].$$

By part (a), $P^n(E_1) \geq 1 - \varepsilon/2$. Then, event E_2 is defined the same way is in the proof of part (b), $P(E_1 \cap E_2) > 1 - \varepsilon$ by a union bound argument, and combining the inequalities in E_1 and E_2 produces the statement of part (c). ■

Proof of Theorem 1.4.2.

Part (a). Let E_1 denote the event that, for all $\rho \in \mathscr{P}_\pi(\Theta)$ simultaneously it holds that

$$\int_\Theta R(\theta) d\rho(\theta) \leq \int_\Theta R_n(\theta) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho,\pi) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right]. \tag{1.67}$$

Let E_2 denote the event that, for all $\rho \in \mathscr{P}_\pi(\Theta)$ simultaneously it holds that

$$\int_\Theta R_n(\theta)d\rho(\theta) \leq \int_\Theta R(\theta)d\rho(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right]. \tag{1.68}$$

Lastly, let $u^* = u^*(B, \lambda/2)$ as specified in Definition 1.3.3 and let E_3 denote the event that

$$\int_\Theta K_n(\theta)d\rho^*_{\lambda/2,u^*}(\theta) - \int_\Theta K(\theta)d\rho^*_{\lambda/2,u^*}(\theta) \leq \sqrt{\frac{M_c^2 \log\frac{3}{\varepsilon}}{2n\kappa^2}}, \tag{1.69}$$

where $\rho^*_{\lambda/2,u}$ is given in Definition 1.3.2.

By Theorem 1.4.1 (a), applied to each $s \in \{-1,1\}$ with $V_n(\theta) = R_n(\theta),\ V(\theta) = R(\theta)$, and by Lemma 1.A.4, respectively, we have

$$P^n(E_1) \geq 1 - \frac{\varepsilon}{3},\ P^n(E_2) \geq 1 - \frac{\varepsilon}{3},\ \text{and}\ P^n(E_3) \geq 1 - \frac{\varepsilon}{3}.$$

Applying a union bound argument as in the proof of Theorem 1.4.1 (b), it holds that $P^n(E_1 \cap E_2 \cap E_3) \geq 1 - \varepsilon$. From the remainder of the proof, we work assuming the intersection of these three events. We show the event in Theorem 1.4.2 (a) is implied by their intersection, hence the event of interest contains this intersection and has probability greater than or equal $1 - \varepsilon$.

We consider two possible scenarios in conjuncture with events E_1, E_2, and E_3. In the first scenario, suppose that

$$\int_\Theta K_n(\theta)d\rho^*_{\lambda/2,u^*}(\theta) \leq B. \tag{1.70}$$

In this case $\rho^*_{\lambda/2,u^*} \in \widehat{\mathscr{E}}_B$, where $\widehat{\mathscr{E}}_B$ is given by (1.20). Starting from (1.67) with $\rho = \hat{\rho}_{\lambda,\hat{u}}$,

$$\begin{aligned}\int_\Theta R(\theta)d\hat{\rho}_{\lambda,\hat{u}}(\theta) &\leq \int_\Theta R_n(\theta)d\hat{\rho}_{\lambda,\hat{u}}(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\hat{\rho}_{\lambda,\hat{u}},\pi) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right] \\ &= \min_{\rho\in\widehat{\mathscr{E}}_B}\left\{\int_\Theta R_n(\theta)d\rho(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right\} + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right] \\ &\leq \int_\Theta R_n(\theta)d\rho^*_{\lambda/2,u^*}(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho^*_{\lambda/2,u^*},\pi) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right].\end{aligned}$$

The second equality above follows from Lemma 1.3.2 (a). Now, consider (1.68) with $\rho = \rho^*_{\lambda/2,u^*}$. Plugging this inequality into the right-hand side of the above inequality produces

$$\begin{aligned}\int_\Theta R(\theta)d\hat{\rho}_{\lambda,\hat{u}}(\theta) &\leq \int_\Theta R(\theta)d\rho^*_{\lambda/2,u^*}(\theta) + \frac{2}{\lambda}D_{\mathrm{KL}}(\rho^*_{\lambda/2,u^*},\pi) + \frac{2}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right] \\ &= \min_{\rho\in\mathscr{E}_B}\left\{\int_\Theta R(\theta)d\rho(\theta) + \frac{2}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right\} + \frac{2}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right] \\ &\leq \min_{\rho\in\mathscr{E}_B}\left\{\int_\Theta R(\theta)d\rho(\theta) + \frac{2}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right\} + \frac{2}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right] + \hat{u}\sqrt{\frac{M_c^2\log\frac{3}{\varepsilon}}{2n\kappa^2}},\end{aligned}$$

where the equality in the second row follows from Lemma 1.A.9 (b) and the final inequality holds as $\hat{u} \geq 0$. Thus the result of part (a) holds in the first scenario described by (1.70), noting that for $\rho \in \mathscr{P}(\Theta)$, $R(f_{G,\rho}) = \int_\Theta R(\theta)d\rho(\theta)$.

In the second and only remaining scenario, we consider when

$$\int_\Theta K_n(\theta)d\rho^*_{\lambda/2,u^*}(\theta) > B. \tag{1.71}$$

If we set

$$B' = \int_\Theta K_n(\theta)d\rho^*_{\lambda/2,u^*}(\theta), \tag{1.72}$$

then it holds that $\rho^*_{\lambda/2,u^*} \in \widehat{\mathscr{E}}_{B'}$. Again starting from the event in (1.67) with $\rho = \hat{\rho}_{\lambda,\hat{u}}$, we obtain

$$
\begin{align*}
&\int_\Theta R(\theta)d\hat{\rho}_{\lambda,\hat{u}}(\theta) \\
&\leq \int_\Theta R_n(\theta)d\hat{\rho}_{\lambda,\hat{u}}(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\hat{\rho}_{\lambda,\hat{u}},\pi) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right] \\
&= \int_\Theta R_n(\theta)d\hat{\rho}_{\lambda,\hat{u}}(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\hat{\rho}_{\lambda,\hat{u}},\pi) + \hat{u}\left(\int_\Theta K_n(\theta)d\hat{\rho}_{\lambda,\hat{u}}(\theta) - B\right) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right] \tag{1.73} \\
&= \int_\Theta R_n(\theta)d\hat{\rho}_{\lambda,\hat{u}}(\theta) + \hat{u}\left(\int_\Theta K_n(\theta)d\hat{\rho}_{\lambda,\hat{u}}(\theta) - B'\right) + \frac{1}{\lambda}D_{\mathrm{KL}}(\hat{\rho}_{\lambda,\hat{u}},\pi) \\
&\quad + \hat{u}\left(\int_\Theta K_n(\theta)d\rho^*_{\lambda/2,u^*}(\theta) - B\right) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right] \tag{1.74} \\
&\leq \sup_{u\geq 0}\left[\int_\Theta R_n(\theta)d\hat{\rho}_{\lambda,u}(\theta) + u\left(\int_\Theta K_n(\theta)d\hat{\rho}_{\lambda,u}(\theta) - B'\right) + \frac{1}{\lambda}D_{\mathrm{KL}}(\hat{\rho}_{\lambda,u},\pi)\right] \\
&\quad + \hat{u}\left(\int_\Theta K_n(\theta)d\rho^*_{\lambda/2,u^*}(\theta) - B\right) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right] \\
&= \min_{\rho\in\widehat{\mathscr{E}}_{B'}}\left\{\int_\Theta R_n(\theta)d\rho(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right\} \\
&\quad + \hat{u}\left(\int_\Theta K_n(\theta)d\rho^*_{\lambda/2,u^*}(\theta) - B\right) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right] \tag{1.75} \\
&\leq \min_{\rho\in\widehat{\mathscr{E}}_{B'}}\left\{\int_\Theta R_n(\theta)d\rho(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right\} + \hat{u}\sqrt{\frac{M_c^2\log\frac{3}{\varepsilon}}{2n\kappa^2}} + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right] \tag{1.76} \\
&\leq \int_\Theta R_n(\theta)d\rho^*_{\lambda/2,u^*}(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho^*_{\lambda/2,u^*},\pi) + \hat{u}\sqrt{\frac{M_c^2\log\frac{3}{\varepsilon}}{2n\kappa^2}} + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right] \tag{1.77} \\
&\leq \int_\Theta R(\theta)d\rho^*_{\lambda/2,u^*}(\theta) + \frac{2}{\lambda}D_{\mathrm{KL}}(\rho^*_{\lambda/2,u^*},\pi) + \hat{u}\sqrt{\frac{M_c^2\log\frac{3}{\varepsilon}}{2n\kappa^2}} + \frac{2}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right] \tag{1.78} \\
&= \min_{\rho\in\mathscr{E}_B}\left\{\int_\Theta R(\theta)d\rho(\theta) + \frac{2}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right\} + \frac{2}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{3}{\varepsilon}\right] + \hat{u}\sqrt{\frac{M_c^2\log\frac{3}{\varepsilon}}{2n\kappa^2}}. \tag{1.79}
\end{align*}
$$

In the above, step (1.73) follows from the properties of $\hat{u} = \hat{u}(B,\lambda)$ in Lemma 1.3.2 (a). In step (1.74) we simply added and subtracted $\hat{u}B'$ with B' given in (1.72). Step (1.75) follows from

Lemma 1.A.9 (a). Step (1.76) follows from (1.69) and the observation that $\int_\Theta K(\theta) d\rho^*_{\lambda/2,u^*}(\theta)$ is always less than or equal to B by Lemma 1.A.9 (b). Step (1.77) follows from fact that $\rho^*_{\lambda/2,u^*} \in \widehat{\mathscr{E}}_{B'}$ by the construction of B' in (1.72). (1.78) follows from (1.68) with $\rho = \rho^*_{\lambda/2,u^*}$ and lastly (1.79) follows from Lemma 1.A.9 (b).

It follows that the result in part (a) also holds in the second scenario in (1.71) which completes the proof for this part.

Part (b). Now, let E_1 denote the event that, for all $\rho \in \mathscr{P}_\pi(\Theta)$ simultaneously it holds that

$$\int_\Theta R(\theta) d\rho(\theta) \leq \int_\Theta R_n(\theta) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{4}{\varepsilon}\right]. \tag{1.80}$$

Let E_2 denote the event that

$$\begin{aligned}
&\int_\Theta R_n(\theta) d\hat{\rho}_{\lambda,u}(\theta) + u\int_\Theta K_n(\theta) d\hat{\rho}_{\lambda,u}(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}\left(\hat{\rho}_{\lambda,u}, \pi\right) \\
&\leq \int_\Theta R(\theta) d\rho^*_{\lambda,u}(\theta) + u\int_\Theta K(\theta) d\rho^*_{\lambda,u}(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho^*_{\lambda,u}, \pi) + \sqrt{\frac{(M_y + uM_c)^2 \log(4/\varepsilon)}{2n\kappa^2}},
\end{aligned} \tag{1.81}$$

and let E_3 denote the event that

$$\begin{aligned}
&\int_\Theta K(\theta) d\hat{\rho}_{\lambda,u}(\theta) - \int_\Theta K_n(\theta) d\hat{\rho}_{\lambda,u}(\theta) \\
&\leq \frac{\sqrt{2}\,(M_y + uM_c)}{\kappa\sqrt{n}} \sqrt{\log\left(2\sqrt{n}\right) + \log\frac{4}{\varepsilon}} + \frac{\lambda\,(M_y + uM_c)^2}{2n\kappa^2} + \frac{1}{\lambda}\left[\frac{\lambda^2 M_c^2}{8n\kappa^2} + \log\frac{4}{\varepsilon}\right] \\
&= U_1\left(\varepsilon; \lambda, u, n\right) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_c^2}{8n\kappa^2} + \log\frac{4}{\varepsilon}\right]
\end{aligned} \tag{1.82}$$

By Theorem 1.4.1 (a), applied with $s = -1$, $V_n(\theta) = R_n(\theta)$, $V(\theta) = R(\theta)$, and $M_\ell = M_y$, we have that $P^n(E_1) = \varepsilon/4$. By Lemma 1.A.7, $P^n(E_2) = \varepsilon/4$. And lastly, by Theorem 1.4.1 (c) with $V_n(\theta) = K_n(\theta)$, $V(\theta) = K(\theta)$, and $M_\ell = M_c$, it holds that $P^n(E_3) = \varepsilon/2$. Again applying a union bound argument similar to that in the proof of Theorem 1.4.1 (b), we have

$$P^n\left(E_1 \cap E_2 \cap E_3\right) \geq 1-\varepsilon.$$

As in part (a), we prove the result by showing that the intersection of these events implies the event in the result.

Recall,

$$B\left(\hat{\rho}_{\lambda,u}\right) = \int_{\Theta} K(\theta)d\hat{\rho}_{\lambda,u}(\theta) \text{ and } \widehat{B}\left(\hat{\rho}_{\lambda,u}\right) = \int_{\Theta} K_n(\theta)d\hat{\rho}_{\lambda,u}(\theta).$$

Then, the event E_3 described in (1.82) can be stated

$$B\left(\hat{\rho}_{\lambda,u}\right) - \widehat{B}\left(\hat{\rho}_{\lambda,u}\right) \leq U_1\left(\varepsilon;\lambda,u,n\right) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_c^2}{8n\kappa^2} + \log\frac{4}{\varepsilon}\right]. \tag{1.83}$$

Now, starting from (1.80) with $\rho = \hat{\rho}_{\lambda,u}$,

$$\begin{align*}
&\int_\Theta R(\theta)d\hat{\rho}_{\lambda,u}(\theta) \\
&\leq \int_\Theta R_n(\theta)d\hat{\rho}_{\lambda,u}(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\hat{\rho}_{\lambda,u},\pi) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{4}{\varepsilon}\right] \\
&= \int_\Theta R_n(\theta)d\hat{\rho}_{\lambda,u}(\theta) + u\left(\int_\Theta K_n(\theta)d\hat{\rho}_{\lambda,u} - \widehat{B}\left(\hat{\rho}_{\lambda,u}\right)\right) + \frac{1}{\lambda}D_{\mathrm{KL}}(\hat{\rho}_{\lambda,u},\pi) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{4}{\varepsilon}\right] \\
&= \int_\Theta R_n(\theta)d\hat{\rho}_{\lambda,u}(\theta) + u\int_\Theta K_n(\theta)d\hat{\rho}_{\lambda,u} + \frac{1}{\lambda}D_{\mathrm{KL}}(\hat{\rho}_{\lambda,u},\pi) \\
&\qquad - uB\left(\hat{\rho}_{\lambda,u}\right) + u\left(B\left(\hat{\rho}_{\lambda,u}\right) - \widehat{B}\left(\hat{\rho}_{\lambda,u}\right)\right) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{4}{\varepsilon}\right] \\
&\leq \int_\Theta R(\theta)d\rho^*_{\lambda,u}(\theta) + u\int_\Theta K(\theta)d\rho^*_{\lambda,u}(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho^*_{\lambda,u},\pi) + \sqrt{\frac{(M_y + uM_c)^2\log(4/\varepsilon)}{2n\kappa^2}} \\
&\qquad - uB\left(\hat{\rho}_{\lambda,u}\right) + u\left(B\left(\hat{\rho}_{\lambda,u}\right) - \widehat{B}\left(\hat{\rho}_{\lambda,u}\right)\right) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{4}{\varepsilon}\right] \tag{1.84} \\
&= \int_\Theta R(\theta)d\rho^*_{\lambda,u}(\theta) + u\left(\int_\Theta K(\theta)d\rho^*_{\lambda,u}(\theta) - B\left(\hat{\rho}_{\lambda,u}\right)\right) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho^*_{\lambda,u},\pi) + \sqrt{\frac{(M_y + uM_c)^2\log(4/\varepsilon)}{2n\kappa^2}} \\
&\qquad + u\left(B\left(\hat{\rho}_{\lambda,u}\right) - \widehat{B}\left(\hat{\rho}_{\lambda,u}\right)\right) + \frac{1}{\lambda}\left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log\frac{4}{\varepsilon}\right] \\
&\leq \int_\Theta R(\theta)d\rho^*_{\lambda,u}(\theta) + u\left(\int_\Theta K(\theta)d\rho^*_{\lambda,u}(\theta) - B\left(\hat{\rho}_{\lambda,u}\right)\right) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho^*_{\lambda,u},\pi) + \sqrt{\frac{(M_y + uM_c)^2\log(4/\varepsilon)}{2n\kappa^2}} \\
&\qquad + uU_1\left(\varepsilon;\lambda,u,n\right) + \frac{1}{\lambda}\left[\frac{\lambda^2\left(M_y^2 + uM_c^2\right)}{8n\kappa^2} + (1+u)\log\frac{4}{\varepsilon}\right] \tag{1.85} \\
&\leq \sup_{a\geq 0}\left[\int_\Theta R(\theta)d\rho^*_{\lambda,a}(\theta) + a\left(\int_\Theta K(\theta)d\rho^*_{\lambda,a}(\theta) - B\left(\hat{\rho}_{\lambda,u}\right)\right) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho^*_{\lambda,a},\pi)\right] \\
&\qquad + \sqrt{\frac{(M_y + uM_c)^2\log(4/\varepsilon)}{2n\kappa^2}} + uU_1\left(\varepsilon;\lambda,u,n\right) + \frac{1}{\lambda}\left[\frac{\lambda^2\left(M_y^2 + uM_c^2\right)}{8n\kappa^2} + (1+u)\log\frac{4}{\varepsilon}\right] \\
&= \min_{\rho\in\mathscr{E}_{B(\hat{\rho}_{\lambda,u})}}\left\{\int_\Theta R(\theta)d\rho(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right\} \\
&\qquad\qquad + \sqrt{\frac{(M_y + uM_c)^2\log(4/\varepsilon)}{2n\kappa^2}} + uU_1\left(\varepsilon;\lambda,u,n\right) + \frac{1}{\lambda}\left[\frac{\lambda^2\left(M_y^2 + uM_c^2\right)}{8n\kappa^2} + (1+u)\log\frac{4}{\varepsilon}\right]. \tag{1.86}
\end{align*}$$

In the above, (1.84) follows from plugging in (1.81), (1.85) follows from plugging in (1.83), and lastly (1.86) follows from Lemma 1.A.9 (c). Switching the notation to $\int_\Theta R(\theta)d\rho(\theta) = R(f_{G,\rho})$

for $\rho \in \mathscr{P}(\Theta)$ and utilizing the definition of $U_2(\varepsilon;\lambda,u,n)$, the above yields the statement in part (b) of the Theorem. ■

Proofs for Subsection 1.4.2: Normal Prior

Proof of Theorem 1.4.3. We will use the following properties in the proofs of part (a) and (b). For treatment assignment rules of the form in (1.10), when $\|\theta\| \neq 0$, it holds that $f_\theta(x) = f_{\theta/\|\theta\|}(x)$ for all $x \in \mathscr{X}$. As we are presuming that $\overline{\theta} \neq 0$ and $\overline{\theta}_u \neq 0$ (almost surely), with probability one we can find a values $\overline{\theta}$ and $\overline{\theta}_u$ such that $\|\overline{\theta}\| = 1$ and $\|\overline{\theta}_u\| = 1$. We assume $\overline{\theta}$ and $\overline{\theta}_u$ are selected to have this property for the remainder of the proof. Below, for integration over $\Theta = \mathbb{R}^q$, we write $\int \ldots$ in place of $\int_{\mathbb{R}^q} \ldots$

Observe that for $\theta, \theta_1 \in \mathbb{R}^q$ such that $\|\theta_1\| = 1$ and $\|\theta\| \neq 0$,

$$\begin{aligned}
R(\theta) - R(\theta_1) &= W(f_\theta) - W(f_{\theta_1}) && (1.87)\\
&= E_Q\left[(Y_1 - Y_0)\left(f_\theta(X) - f_{\theta_1}(X)\right)\right] \\
&\leq M_y E_P\left[\left|1\{\phi(X)^\intercal\theta > 0\} - 1\{\phi(X)^\intercal\theta_1 > 0\}\right|\right] && (1.88)\\
&= M_y P\left[(\phi(X)^\intercal\theta)(\phi(X)^\intercal\theta_1) < 0\right] \\
&= M_y P\left[\left(\phi(X)^\intercal\frac{\theta}{\|\theta\|}\right)(\phi(X)^\intercal\theta_1) < 0\right] \\
&\leq M_y \nu \left\|\frac{\theta}{\|\theta\|} - \theta_1\right\| && (1.89)\\
&\leq M_y 2\nu \left\|\theta - \theta_1\right\|, && (1.90)
\end{aligned}$$

where (1.87) follows from the definition of welfare regret, (1.88) follows from Assumption 1.3.1 (iii) and the fact that the distribution of X is determined by P as well as Q, (1.89) follows from Assumption 1.4.3, and (1.90) follows from the fact that with θ, θ_1 as above,

$$\left\|\frac{\theta}{\|\theta\|} - \theta_1\right\| \leq \|\theta - \theta_1\|.$$

As a consequence of (1.90), for any $\sigma > 0$,

$$\begin{aligned}\int R(\theta) d\Phi_{\theta_1,\sigma^2}(\theta) &= R(\theta_1) + \int [R(\theta) - R(\theta_1)]\, d\Phi_{\theta_1,\sigma^2}(\theta) \\ &\leq R(\theta_1) + 2M_y \nu \int \|\theta - \theta_1\|\, d\Phi_{\theta_1,\sigma^2}(\theta) \\ &\leq R(\theta_1) + 2M_y \nu\sigma\sqrt{q}, \qquad (1.91)\end{aligned}$$

where we have used the fact that for $\theta \sim \Phi_{\theta_1,\sigma^2}$, $||\theta - \theta_1|| \sim \sigma H^{1/2}$ with $H \sim \chi^2(q)$. Then, by Jensen's inequality, $E\sigma H^{1/2} \leq \sigma(EH)^{1/2} = \sigma(q)^{1/2}$.

Following nearly identical steps, now starting with the definition of the expected costs $K(\theta)$ and $K(\overline{\theta})$, it is straightforward to derive that, for $\theta, \theta_1 \in \mathbb{R}^q$ such that $\|\theta_1\| = 1$ and $\|\theta\| \neq 0$,

$$K(\theta) - K(\theta_1) \leq M_c 2\nu \|\theta - \theta_1\|,$$

and for $\sigma > 0$,

$$\int K(\theta) d\Phi_{\theta_1,\sigma^2}(\theta) \leq K(\theta_1) + 2M_c \nu\sigma\sqrt{q}. \qquad (1.92)$$

Lastly, before considering part (a) and (b) separately, note that by Lemma 1.A.5, with $\sigma_\pi = 1/\sqrt{q}$, $\sigma_\rho = 1/(2\sqrt{nq})$, and $\|\theta_1\| = 1$,

$$D_{\mathrm{KL}}\left(\Phi_{\theta_1,\sigma_\rho^2}, \Phi_{0,\sigma_\pi^2}\right) = \frac{q}{2}\left[\frac{1}{4n} + \log(4n)\right]. \qquad (1.93)$$

Part (a). We consider the posterior distribution $\widetilde{\rho} = \Phi_{\overline{\theta},\sigma_\rho^2}$ with $\sigma_\rho = 1/(2\sqrt{nq})$ so that $D_{\mathrm{KL}}(\widetilde{\rho}, \pi)$ is given by (1.93). Next, define

$$B' = B + \frac{\nu M_c}{\sqrt{n}}. \qquad (1.94)$$

Assumptions 1.3.2 and 1.3.3 are met and Assumptions 1.4.3 and 1.3.4 are assumed to hold so we can apply Theorem 1.4.2 (a). Starting from there, with probability at least $1 - \varepsilon$ we have

$$\begin{aligned}
&\int R(\theta)d\hat{\rho}_{\lambda,\hat{u}}(\theta)\\
&\leq \min_{\rho\in\mathscr{E}_B}\left\{\int R(\theta)d\rho(\theta)+\frac{2}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right\}+\frac{2}{\lambda}\left[\frac{\lambda^2M_y^2}{8n\kappa^2}+\log\frac{3}{\varepsilon}\right]+\hat{u}\sqrt{\frac{M_c^2\log\frac{3}{\varepsilon}}{2n\kappa^2}}\\
&=\int R(\theta)d\rho^*_{\lambda/2,u^*}(\theta)+\frac{2}{\lambda}D_{\mathrm{KL}}(\rho^*_{\lambda/2,u^*},\pi)+\frac{2}{\lambda}\left[\frac{\lambda^2M_y^2}{8n\kappa^2}+\log\frac{3}{\varepsilon}\right]+\hat{u}\sqrt{\frac{M_c^2\log\frac{3}{\varepsilon}}{2n\kappa^2}} \qquad (1.95)\\
&=\int R(\theta)d\rho^*_{\lambda/2,u^*}(\theta)+u^*\left(\int K(\theta)d\rho^*_{\lambda/2,u^*}(\theta)-B\right)+\frac{2}{\lambda}D_{\mathrm{KL}}(\rho^*_{\lambda/2,u^*},\pi)\\
&\quad+\frac{2}{\lambda}\left[\frac{\lambda^2M_y^2}{8n\kappa^2}+\log\frac{3}{\varepsilon}\right]+\hat{u}\sqrt{\frac{M_c^2\log\frac{3}{\varepsilon}}{2n\kappa^2}} \qquad (1.96)\\
&=\int R(\theta)d\rho^*_{\lambda/2,u^*}(\theta)+u^*\left(\int K(\theta)d\rho^*_{\lambda/2,u^*}(\theta)-B'\right)+\frac{2}{\lambda}D_{\mathrm{KL}}(\rho^*_{\lambda/2,u^*},\pi)\\
&\quad+u^*\left(B'-B\right)+\frac{2}{\lambda}\left[\frac{\lambda^2M_y^2}{8n\kappa^2}+\log\frac{3}{\varepsilon}\right]+\hat{u}\sqrt{\frac{M_c^2\log\frac{3}{\varepsilon}}{2n\kappa^2}}\\
&\leq \sup_{u\geq 0}\left[\int R(\theta)d\rho^*_{\lambda/2,u}+u\left(\int K(\theta)d\rho^*_{\lambda,u}(\theta)-B'\right)+\frac{2}{\lambda}D_{\mathrm{KL}}(\rho^*_{\lambda,u},\pi)\right]\\
&\quad+u^*\left(B'-B\right)+\frac{2}{\lambda}\left[\frac{\lambda^2M_y^2}{8n\kappa^2}+\log\frac{3}{\varepsilon}\right]+\hat{u}\sqrt{\frac{M_c^2\log\frac{3}{\varepsilon}}{2n\kappa^2}}\\
&=\min_{\rho\in\mathscr{E}_{B'}}\left\{\int R(\theta)d\rho(\theta)+\frac{2}{\lambda}D_{\mathrm{KL}}\left(\rho,\pi\right)\right\}+u^*\frac{\nu M_c}{\sqrt{n}}+\frac{2}{\lambda}\left[\frac{\lambda^2M_y^2}{8n\kappa^2}+\log\frac{3}{\varepsilon}\right]+\hat{u}\sqrt{\frac{M_c^2\log\frac{3}{\varepsilon}}{2n\kappa^2}}\\
&\qquad (1.97)
\end{aligned}$$

In the above, (1.95) and (1.96) follow from Lemma 1.A.9 (b) while (1.97) follows from applying Lemma 1.A.9 (b) and the definition of B' in (1.94).

From (1.92) with $\overline{\theta}$ in the place of θ_1 and with $\sigma_\rho = 1/(2\sqrt{nq})$, we have

$$\int K(\theta)d\widetilde{\rho}(\theta)=\int K(\theta)d\Phi_{\overline{\theta},\sigma_\rho^2}(\theta)\leq K\left(\overline{\theta}\right)+\frac{\nu M_c}{\sqrt{n}}\leq B' \qquad (1.98)$$

as, by the definition of $\overline{\theta}$, $K(\overline{\theta})\leq B$. Therefore $\widetilde{\rho}\in\mathscr{E}_{B'}$. From (1.97), we have, with probability at least $1-\varepsilon$,

$$
\begin{aligned}
&\int R(\theta) d\hat{\rho}_{\lambda,\hat{u}}(\theta) \\
&\leq \min_{\rho \in \mathscr{E}_{B'}} \left\{ \int R(\theta) d\rho(\theta) + \frac{2}{\lambda} D_{\mathrm{KL}}(\rho, \pi) \right\} + u^* \frac{\nu M_c}{\sqrt{n}} + \frac{2}{\lambda} \left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log \frac{3}{\varepsilon} \right] + \hat{u} \sqrt{\frac{M_c^2 \log \frac{3}{\varepsilon}}{2n\kappa^2}} \\
&\leq \int R(\theta) d\widetilde{\rho}(\theta) + \frac{2}{\lambda} D_{\mathrm{KL}}(\widetilde{\rho}, \pi) + u^* \frac{\nu M_c}{\sqrt{n}} + \frac{2}{\lambda} \left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log \frac{3}{\varepsilon} \right] + \hat{u} \sqrt{\frac{M_c^2 \log \frac{3}{\varepsilon}}{2n\kappa^2}} \\
&\leq R\left(\overline{\theta}\right) + \frac{\nu M_y}{\sqrt{n}} + \frac{q}{\lambda} \left[\frac{1}{4n} + \log(4n) \right] + u^* \frac{\nu M_c}{\sqrt{n}} + \frac{2}{\lambda} \left[\frac{\lambda^2 M_y^2}{8n\kappa^2} + \log \frac{3}{\varepsilon} \right] + \hat{u} \sqrt{\frac{M_c^2 \log \frac{3}{\varepsilon}}{2n\kappa^2}}.
\end{aligned}
$$

In the last step, we have applied the properties in (1.91) and (1.93) with $\overline{\theta}$ taking the role of θ_1. Plugging in $\lambda = \kappa\sqrt{nq}/M_y$ and rearranging terms then produces the result in (a) with

$$
\overline{U}_1(n;q) = \sqrt{\frac{q}{n}} \left[\frac{\nu M_y}{\sqrt{q}} + \frac{M_y}{\kappa} \left(\frac{1}{4} + \frac{1}{4n} \right) \right].
$$

Part (b). As a starting point, we utilize the setup and initial steps of the proof of Theorem 1.4.2 (b). Assume the same the definitions of events E_1, E_2 and E_3 as in (1.80), (1.81), (1.82), respectively. Following that proof up to (1.85), we have that with probability at least $1-\varepsilon$,

$$
\begin{aligned}
&\int R(\theta) d\hat{\rho}_{\lambda,u}(\theta) \\
&\leq \int R(\theta) d\rho^*_{\lambda,u}(\theta) + u \left(\int K(\theta) d\rho^*_{\lambda,u}(\theta) - B\left(\hat{\rho}_{\lambda,u}\right) \right) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho^*_{\lambda,u}, \pi) \\
&\quad + \sqrt{\frac{(M_y + uM_c)^2 \log(4/\varepsilon)}{2n\kappa^2}} + uU_1(\varepsilon; \lambda, u, n) + \frac{1}{\lambda} \left[\frac{\lambda^2 \left(M_y^2 + uM_c^2\right)}{8n\kappa^2} + (1+u) \log \frac{4}{\varepsilon} \right],
\end{aligned}
\tag{1.99}
$$

where $B(\hat{\rho}_{\lambda,u}) = \int K(\theta) d\hat{\rho}_{\lambda,u} = K(f_{G,\hat{\rho}_{\lambda,u}})$ and $U_1(\varepsilon; \lambda, u, n)$ is defined in Theorem 1.4.2 (b).

Now we will consider the posterior $\widetilde{\rho} = \Phi_{\overline{\theta}_u, \sigma_\rho^2}$ with $\sigma_\rho = 1/(2\sqrt{nq})$. Utilizing (1.93) with π as described in the theorem, now with $\overline{\theta}_u$ in place of θ_1, with probability one we have

$$D_{\mathrm{KL}}(\widetilde{\rho}, \pi) = \frac{q}{2}\left[\frac{1}{4n} + \log(4n)\right]. \tag{1.100}$$

Additionally, we now define

$$B' = B\left(\hat{\rho}_{\lambda,u}\right) + \frac{\nu M_c}{\sqrt{n}}. \tag{1.101}$$

From (1.92) with $\overline{\theta}_u$ in the place of θ_1 and with $\sigma_\rho = 1/(2\sqrt{nq})$, with probability one we have

$$\int K(\theta)d\widetilde{\rho}(\theta) = \int K(\theta)d\Phi_{\overline{\theta}_u,\sigma_\rho^2}(\theta) \leq K\left(\overline{\theta}_u\right) + \frac{\nu M_c}{\sqrt{n}} \leq B', \tag{1.102}$$

because, by the definition of $\overline{\theta}_u$ we have $K(\overline{\theta}_u) \leq B\left(\hat{\rho}_{\lambda,u}\right)$ (a.s.). It follows that with probability one, $\widetilde{\rho} \in \mathscr{E}_{B'}$.

Returning to (1.99), we have, with probability at least $1-\varepsilon$,

$$
\begin{aligned}
&\int R(\theta) d\hat{\rho}_{\lambda,u}(\theta) \\
&\leq \int R(\theta) d\rho^*_{\lambda,u}(\theta) + u\left(\int K(\theta) d\rho^*_{\lambda,u}(\theta) - B\left(\hat{\rho}_{\lambda,u}\right)\right) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho^*_{\lambda,u}, \pi) + \sqrt{\frac{(M_y + uM_c)^2 \log(4/\varepsilon)}{2n\kappa^2}} \\
&\quad + uU_1\left(\varepsilon; \lambda, u, n\right) + \frac{1}{\lambda}\left[\frac{\lambda^2\left(M_y^2 + uM_c^2\right)}{8n\kappa^2} + (1+u)\log\frac{4}{\varepsilon}\right] \\
&\leq \int R(\theta) d\rho^*_{\lambda,u}(\theta) + u\left(\int K(\theta) d\rho^*_{\lambda,u}(\theta) - B'\right) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho^*_{\lambda,u}, \pi) + \sqrt{\frac{(M_y + uM_c)^2 \log(4/\varepsilon)}{2n\kappa^2}} \\
&\quad + u\left(B' - B\left(\hat{\rho}_{\lambda,u}\right)\right) + uU_1\left(\varepsilon; \lambda, u, n\right) + \frac{1}{\lambda}\left[\frac{\lambda^2\left(M_y^2 + uM_c^2\right)}{8n\kappa^2} + (1+u)\log\frac{4}{\varepsilon}\right] \\
&\leq \sup_{a \geq 0}\left[\int R(\theta) d\rho^*_{\lambda,a}(\theta) + u\left(\int K(\theta) d\rho^*_{\lambda,a}(\theta) - B'\right) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho^*_{\lambda,a}, \pi)\right] \\
&\quad + \sqrt{\frac{(M_y + uM_c)^2 \log(4/\varepsilon)}{2n\kappa^2}} + u\left(\frac{\nu M_c}{\sqrt{n}}\right) + uU_1\left(\varepsilon; \lambda, u, n\right) + \frac{1}{\lambda}\left[\frac{\lambda^2\left(M_y^2 + uM_c^2\right)}{8n\kappa^2} + (1+u)\log\frac{4}{\varepsilon}\right]
\end{aligned}
\tag{1.103}
$$

$$
\begin{aligned}
&= \inf_{\rho \in \mathscr{E}_{B'}}\left[\int R(\theta) d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}\left(\rho, \pi\right)\right] + \sqrt{\frac{(M_y + uM_c)^2 \log(4/\varepsilon)}{2n\kappa^2}} \\
&\quad + u\left(\frac{\nu M_c}{\sqrt{n}}\right) + uU_1\left(\varepsilon; \lambda, u, n\right) + \frac{1}{\lambda}\left[\frac{\lambda^2\left(M_y^2 + uM_c^2\right)}{8n\kappa^2} + (1+u)\log\frac{4}{\varepsilon}\right]
\end{aligned}
\tag{1.104}
$$

$$
\begin{aligned}
&\leq \int R(\theta) d\widetilde{\rho}(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}\left(\widetilde{\rho}, \pi\right) + \sqrt{\frac{(M_y + uM_c)^2 \log(4/\varepsilon)}{2n\kappa^2}} \\
&\quad + u\left(\frac{\nu M_c}{\sqrt{n}}\right) + uU_1\left(\varepsilon; \lambda, u, n\right) + \frac{1}{\lambda}\left[\frac{\lambda^2\left(M_y^2 + uM_c^2\right)}{8n\kappa^2} + (1+u)\log\frac{4}{\varepsilon}\right]
\end{aligned}
\tag{1.105}
$$

$$
\begin{aligned}
&\leq R\left(\overline{\theta}_u\right) + \frac{\nu M_y}{\sqrt{n}} + \frac{q}{2\lambda}\left[\frac{1}{4n} + \log\left(4n\right)\right] + \sqrt{\frac{(M_y + uM_c)^2 \log(4/\varepsilon)}{2n\kappa^2}} \\
&\quad + u\left(\frac{\nu M_c}{\sqrt{n}}\right) + uU_1\left(\varepsilon; \lambda, u, n\right) + \frac{1}{\lambda}\left[\frac{\lambda^2\left(M_y^2 + uM_c^2\right)}{8n\kappa^2} + (1+u)\log\frac{4}{\varepsilon}\right]
\end{aligned}
\tag{1.106}
$$

Above, (1.103) follows from (1.101) and the fact that the supremum there is greater than or equal to the object it replaces, (1.104) follows from Lemma 1.A.9 (c), (1.105) follows from having $\widetilde{\rho} \in \mathscr{E}_{B'}$ with probability one, and lastly (1.106) follows from (1.91), with $\overline{\theta}_u$ in place of θ_1 and $\sigma_\rho = 1/(2\sqrt{nq})$ in place of σ, and utilizing (1.100). Plugging in λ as given in part (b), straightforward manipulations of the expression in (1.106) show that the inequality can be written

$$R\left(f_{G,\hat{\rho}_{\lambda,u}}\right) \leq R\left(\overline{\theta}_u\right) + \frac{M_y+uM_c}{\kappa}\left[\overline{U}_2(n;q,u,\varepsilon)+\overline{U}_3(n;q,u,\varepsilon)+\overline{U}_4(n;q,u)\right],$$

where

$$\overline{U}_2(n;q,u,\varepsilon) = \frac{\sqrt{q}\log\left(2\sqrt{n}\right)+\sqrt{2}u\sqrt{\log\left(2\sqrt{n}\right)+\log\frac{4}{\varepsilon}}}{\sqrt{n}} = \mathscr{O}\left(\frac{\log n}{\sqrt{n}}\right),$$

$$\overline{U}_3(n;q,u,\varepsilon) = \frac{\sqrt{\frac{\log(4/\varepsilon)}{2}}+\frac{1}{\sqrt{q}}(1+u)\log\frac{4}{\varepsilon}}{\sqrt{n}} = \mathscr{O}\left(\frac{1}{\sqrt{n}}\right),$$

and

$$\overline{U}_4(n;q,u) = \frac{\kappa\nu+\sqrt{q}\left(\frac{1}{8n}+\frac{u}{2}\right)}{\sqrt{n}} + \sqrt{\frac{q}{n}}\left(\frac{M_y^2+uM_c^2}{8\left(M_y+uM_c\right)^2}\right) = \mathscr{O}\left(\frac{1}{\sqrt{n}}\right).$$

■

Proofs for Subsection 1.4.3: The Majority Vote Treatment Rule

Proof of Theorem 1.4.4. First, note that

$$f_{\text{mv},\rho}(x) = 1\left\{\int_\Theta f_\theta(x)d\rho(\theta) > \frac{1}{2}\right\} \leq 2\int_\Theta f_\theta(x)d\rho(\theta) = 2f_{G,\rho}(x). \tag{1.107}$$

To see this, note that when $\int_\Theta f(x)d\rho(\theta) \leq 1/2$, $f_{\text{mv},\rho} = 0$ hence the left-hand size of the above inequality is zero while the right-hand size is non-negative and the inequality holds. When $\int_\Theta f_\theta(x)d\rho(\theta) > 1/2$, the left hand side is 1 while the right hand side must be greater than 1, so the inequality holds in all cases.

Next we will show that for any $x \in \mathscr{X}$,

$$\left(\delta_y(x)-\eta_{B(\rho)}\delta_c(x)\right)\left(f^*_{B(\rho)}(x)-f_{\text{mv},\rho}(x)\right) \leq 2\left(\delta_y(x)-\eta_{B(\rho)}\delta_c(x)\right)\left(f^*_{B(\rho)}(x)-f_{G,\rho}(x)\right). \tag{1.108}$$

To see this, first consider $x \in \mathscr{X}$ such that $f^*_{B(\rho)}(x) = 1\{\delta_y(x)-\eta_{B(\rho)}\delta_c(x) > 0\} = 0$. In this case, $\delta_y(x) - \eta_{B(\rho)}\delta_c(x) \leq 0$ and we have

$$\begin{aligned}
\left(\delta_y(x) - \eta_{B(\rho)}\delta_c(x)\right)\left(f^*_{B(\rho)}(x) - f_{\text{mv},\rho}(x)\right) &= \left|\delta_y(x) - \eta_{B(\rho)}\delta_c(x)\right| f_{\text{mv},\rho}(x) \\
&\leq 2\left|\delta_y(x) - \eta_{B(\rho)}\delta_c(x)\right| f_{G,\rho}(x) \\
&= 2\left(\delta_y(x) - \eta_{B(\rho)}\delta_c(x)\right)\left(f^*_{B(\rho)}(x) - f_{G,\rho}(x)\right),
\end{aligned}$$

where the inequality follows from (1.107). To verify (1.108), we now need to check that it holds for $x \in \mathscr{X}$ such that $f^*_{B(\rho)}(x) = 1\{\delta_y(x) - \eta_{B(\rho)}\delta_c(x) > 0\} = 1$. In this case, $\delta_y(x) - \eta_{B(\rho)}\delta_c(x) > 0$, so (1.108) reduces to

$$\left(f^*_{B(\rho)}(x) - f_{\text{mv},\rho}(x)\right) \leq 2\left(f^*_{B(\rho)}(x) - f_{G,\rho}(x)\right). \tag{1.109}$$

First consider $x \in \mathscr{X}$ such that $f_{\text{mv},\rho}(x) = 1$. Then the left-hand size is zero while the right hand side is non-negative as $f_{G,\rho} \in [0,1]$ and $f^*_{B(\rho)(x)} = 1$ in the current assumed setting, so the condition holds. Lastly, if $f_{\text{mv},\rho}(x) = 0$, so that $\int_\Theta f_\theta(x)d\rho(\theta) \leq 1/2$, in the current setting with $f^*_{B(\rho)}(x) = 1$ we then have

$$\begin{aligned}
2\left(f^*_{B(\rho)}(x) - f_{G,\rho}(x)\right) &= 2\left(1 - \int_\Theta f_\theta(x)d\rho(\theta)\right) \\
&\geq 2\left(1 - \frac{1}{2}\right) \\
&= 1 \\
&= f^*_{B(\rho)}(x) - f_{\text{mv},\rho}(x).
\end{aligned}$$

Hence (1.108) holds for all $x \in \mathscr{X}$. Taking the expectation of both sides of that inequality with respect to a draw of X from Q then yields that

$$L_{B(\rho)}\left(f_{\text{mv},\rho}\right) \leq 2L_{B(\rho)}\left(f_{G,\rho}\right). \tag{1.110}$$

To complete the proof, we need to verify that

$$L_{B(\rho)}\left(f_{G,\rho}\right) = R_{B(\rho)}\left(f_{G,\rho}\right). \tag{1.111}$$

Now there are two possibilities to consider. The first is when $\eta_{B(\rho)} = 0$. In this case, we have

$$\begin{aligned} L_{B(\rho)}\left(f_{G,\rho}\right) &= E_Q\left[\delta_y(X)\left(f^*_{B(\rho)} - f_{G,\rho}\right)\right] \\ &= W\left(f^*_{B(\rho)}\right) - W\left(f_{G,\rho}\right) = R_{B(\rho)}\left(f_{G,\rho}\right). \end{aligned}$$

And in the only remaining case, when $\eta_{B(\rho)} > 0$, we also have $K(f^*_{B(\rho)}) = B(\rho)$ by Theorem 1.3.1. As, by the definition of $B(\rho)$, it also holds that $K(f_{G,\rho}) = B(\rho)$, we have

$$\begin{aligned} L_{B(\rho)}\left(f_{G,\rho}\right) &= E_Q\left[\delta_y(X)\left(f^*_{B(\rho)} - f_{G,\rho}\right)\right] - \eta_{B(\rho)} E_Q\left[\delta_c(X)\left(f^*_{B(\rho)} - f_{G,\rho}\right)\right] \\ &= E_Q\left[\delta_y(X)\left(f^*_{B(\rho)} - f_{G,\rho}\right)\right] - \eta_{B(\rho)}\left[K\left(f^*_{B(\rho)}\right) - K\left(f_{G,\rho}\right)\right] \\ &= E_Q\left[\delta_y(X)\left(f^*_{B(\rho)} - f_{G,\rho}\right)\right] \\ &= R_{B(\rho)}\left(f_{G,\rho}\right). \end{aligned}$$

Hence (1.111) holds and combined with (1.110) this completes the proof. ■

Chapter 2

Binary Forecast and Decision Rules via PAC-Bayesian Model Aggregation

Abstract

We consider a PAC-Bayesian model aggregation approach to binary decision or forecast rules when different decision-outcome pairs may have asymmetric payoffs that can vary with observed covariates. The approach estimates a probability distribution over a class of models from which majority vote or stochastic decision rules can be derived. Adopting a utility-based measure of loss considered in Granger and Machina (2006), we show the PAC-Bayesian methodology is well suited to this setting. Non-asymptotic training sample bounds and oracle inequalities familiar in form to counterparts from the 0/1-loss literature are derived for the utility-based setting. The decision rules perform competitively in simulation experiments, achieving higher expected utility than several methods proposed in recent literature. The approach is also well suited to data-rich modeling environments; a constrained version of the learning algorithm produces utility-oriented decision rules with similarities to support vector machines.

2.1 Introduction

Forecasting an uncertain binary outcome arises in a variety of economic decision-making problems. Predicting whether or not a loan will be repaid or which direction an asset price will move are examples where a decision maker's action may vary in tandem with a binary forecast. In general, a decision maker may incur costs or benefits that vary depending on the prediction-outcome pair when making a decision such as to grant or decline a loan. Additionally, payoffs may vary with covariates observed prior to realizing the outcome of interest and these covariates may also influence the likelihood of the outcome. For example, as noted in Elliott and Lieli (2013), development finance institutions may view failing to grant a loan that would be repaid as being more costly when the entity is deemed beneficial to a vulnerable population. At the same time, observable characteristics that quantify this need could be correlated with whether or not a loan will be repaid.

It is well known that there are many successful classification algorithms suitable to a variety of applications. However, asymmetric loss can be a crucial element to decision making and most popular classifiers are not designed around this feature. Maximizing a likelihood function or minimizing a zero-one loss function, or a convex surrogate, does not typically weigh the relative costs of false negatives and false positives according to the preferences of the decision maker. Recently, Elliott and Lieli (2013) proposed a maximum-utility approach. Given a class of parametric decision rules, $\{a(x,\theta) : \mathbb{R}^d \to \{-1,1\},\ \theta \in \Theta\}$, which map covariates $X \in \mathbb{R}^d$ to a binary decision or forecast, the parameters $\hat{\theta} \in \Theta$ are selected as those that maximize the empirical expected utility of the decision maker. Here the binary action or forecast a is associated with an uncertain outcome Y with aligned categories $\{-1,1\}$. The utility maximization framework will be the starting point for our analysis.

There is not a large econometric literature geared at this setting for data-rich environments. First, we point out some recent developments. Su (2020) notes that the trade-off between model class complexity and the propensity to over-fit carries through from empirical risk minimization to the utility maximization setting. He situates the maximum-utility problem in the structural risk minimization paradigm of Vapnik (1982). Building on the analysis of Bartlett et al. (2002), Koltchinskii (2001), and others, he considers a hierarchy of potential model classes with increasing complexity and derives distribution-free and data-driven penalties to select an appropriate model class and decision rule. Another approach was recently considered by Babii et al. (2020) who replace non-convex objects that arise in the utility-maximization problem with convex surrogates.

Here we approach utility-based decision rules from the PAC-Bayesian framework. For a collection (or collections) of decision models associated with a measurable parameter space, this will entail estimating a probability distribution over the model parameters. Then decisions are made by aggregating over all possible decision rules, placing the greatest weight on subsets of the parameter or model space associated with the lowest empirical loss. We adopt a utility-based measure of loss considered in Granger and Machina (2006). Several prior works consider binary classification from the PAC-Bayesian point of view including McAllester (1999b), Langford and Shawe-Taylor (2003), McAllester (2003b), Catoni (2007), Germain et al. (2015), and others. We build in particular on the work of Catoni (2007), Germain et al. (2009), and Alquier et al. (2016). When the utility function is bounded, a lemma of Maurer (2004) enables several key steps of the analysis to proceed as one would in the 0/1 loss setting. This trick is also noted in Germain et al. (2015). In the non-bounded case, our setting turns out to be well suited to higher level assumptions like those in Alquier et al. (2016), where the PAC-Bayesian analysis allows for more general loss functions.

Although estimating a probability measure over a parameter space to form decision rules may seem unfamiliar, it is possible to view a variety of decision rules or classifiers in this light. For example, given covariates $X \in \mathbb{R}^d$, a set of transformations $\phi_j(X) : \mathbb{R}^d \to \mathbb{R}$ for $j = 1, \ldots, M$, and some estimated parameter vector $\hat{\theta} \in \Theta = \mathbb{R}^M$, consider predicting $Y \in \{-1, 1\}$ with

$$\hat{Y} = \operatorname{sign}\left[\sum_{j=1}^{M} \phi_j(X)\hat{\theta}_j\right].$$

The estimated parameter vector $\hat{\theta}$ could come from the method of support vector machine (SVM) or the MU procedure of Elliott and Lieli (2013). Both SVM and MU can result in predictions of the above form. Alternatively, consider the probability distribution $\hat{\rho}(\theta)$ over Θ given by the multivariate normal $N(\hat{\theta}, I_M)$ distribution. In this case, it holds that

$$\operatorname{sign}\left[\int_{\Theta} \operatorname{sign}\left\{\sum_{j=1}^{M} \phi_j(X)\theta_j\right\} d\hat{\rho}(\theta)\right] = \operatorname{sign}\left[\sum_{j=1}^{M} \phi_j(X)\hat{\theta}_j\right], \tag{2.1}$$

as can be seen in Section 2.3.2. The left-hand side above can be interpreted as taking a weighted majority vote over the class of models of the form

$$\operatorname{sign}\left[\sum_{j=1}^{M} \phi_j(X)\theta_j\right], \ \ \theta \in \Theta,$$

where $\hat{\rho}$ determines the weights that different regions of Θ receive. On the other hand, the right-hand side of (2.1) takes the same form as the SVM and MU rules. The PAC-Bayesian approach provides a tractable path to analyzing useful theoretical attributes of decision rules centered around data-dependent distributions $\hat{\rho}$, including those for the SVM and MU methods. This analysis guides the choice of distributions that we focus on in this paper. More broadly, while the PAC-Bayesian framework is useful for deriving competitive learning models (our focus here), this tractable path to analyzing potentially complicated models is a key point of interest itself in the machine learning literature. For example, Neyshabur et al. (2017) derive generalization bounds for deep neural networks in a PAC-Bayesian framework.

There are several attractive characteristics of the PAC-Bayesian approach to utility-oriented decision rules. It allows for a very flexible selection of the decision model class (or classes). Almost any classification model with real parameters can be accommodated. Rather than estimating these parameters by minimizing a 0/1 loss, convex surrogate, or likelihood function, a probability distribution over the parameters that is dependent on a measure of empirical utility is constructed. This puts the greatest weight on regions of the parameter space with high empirical utility and then one can aggregate over potential models in a way that favors these regions. Although the analysis is frequentist in nature, estimation tools from the Bayesian literature such as Markov Chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC) can be applied, sidestepping potential difficulties with computational complexity. By utilizing variational or change-of-measure formulas, the approach allows us to derive training sample bounds that hold with high probability. Additionally, model aggregation can alleviate model misspecification and estimation noise; see, for example, Jiang and Tanner (2008) and Freund et al. (2004) where this is analyzed in different 0/1-loss-based classification settings. Both of these papers utilize exponentially weighted aggregators similar to that employed here. Lastly, as pointed out in Elliott and Lieli (2013), the utility-maximizing decision rule is not unique. It is not unusual in many settings to identify several models with identical or similar in-sample performances but with different out-of-sample performances. Model aggregation makes sense in the context of multiple solutions or multiple near-solutions.

The main contributions of this paper are as follows. We add to the toolbox available for estimating binary choice or forecast rules when the decision maker faces asymmetric payoffs that may depend on the value of observable covariates. The methodology is well suited to data-rich environments and the decision/forecast rules perform very competitively against existing alternatives, exhibiting noticeable gains in expected utility in the simulation environments also studied in Elliott and Lieli (2013) and Su (2020). We develop training sample bounds and oracle inequalities for the decision rules. These are similar in form to existing PAC-Bayesian bounds in alternative settings such as the 0/1 loss which is nested by the utility-based loss adopted here. We show that the theoretical insights, training sample bounds, and modeling guidance of the PAC-Bayesian classification literature can be applied to the utility-oriented setting. Additionally, we illustrate how these concepts and decision rules can be adapted to accommodate the situation with multiple model classes of interest and provide practical guidance regarding implementation. Finally, we try to keep the presentation self-contained and expand on details of the approach. While the PAC-Bayesian methodology has not gained a lot of traction or exposure in the econometric literature, its flexibility and analytical tractability in a variety of machine learning problems suggest that it may prove useful in future econometric applications.

The paper is structured as follows. In Section 2.2 we introduce the decision model and PAC-Bayesian framework. In Section 2.3, we establish the theoretical properties of the PAC-Bayesian decision rules and derive a constrained version of the model, which is easier to implement and provides insight to the PAC-Bayesian machinery in this setting. In Section 2.4 we discuss implementation and estimation, and in Section 2.5 we carry out a simulation study. Section 2.6 concludes. Proofs are given in the appendix.Forecasting an uncertain binary outcome arises in a variety of economic decision-making problems. Predicting whether or not a loan will be repaid or which direction an asset price will move are examples where a decision maker's action may vary in tandem with a binary forecast. In general, a decision maker may incur costs or benefits that vary depending on the prediction-outcome pair when making a decision such as to grant or decline a loan. Additionally, payoffs may vary with covariates observed prior to realizing the outcome of interest and these covariates may also influence the likelihood of the outcome. For example, as noted in Elliott and Lieli (2013), development finance institutions may view failing to grant a loan that would be repaid as being more costly when the entity is deemed beneficial to a vulnerable population. At the same time, observable characteristics that quantify this need could be correlated with whether or not a loan will be repaid.

2.2 Forecasting Framework

2.2.1 Model

To frame the decision problem, we adopt the setting of Elliott and Lieli (2013), tying a binary action or decision to forecasting a binary outcome. This is the standard decision-theoretic framework analyzed in, for example, Granger and Machina (2006). In addition to the discussion below, we refer the reader to Granger and Machina (2006), Elliott and Lieli (2013), and the references therein for further theoretical considerations and additional applications of our setting to problems in economics and other sciences.

The decision maker's problem is to choose an action $a \in \{-1, 1\}$ given an observable vector of covariates $X \in \mathbb{R}^d$ with support $\mathscr{X} \subset \mathbb{R}^d$. The actions are defined in a broad sense and are categorically aligned with a binary outcome variable $Y \in \{-1,1\}$ that is not observable at the time of decision making. Conditional on $X = x$, the outcome variable Y follows a Bernoulli distribution with parameter $P(x)$ where

$$P(x) = \Pr(Y = 1|X = x). \tag{2.2}$$

The payoff or utility function of the decision maker is $U(a,Y,X)$. $U : \{-1,1\} \times \{-1,1\} \times \mathscr{X} \to \mathbb{R}$ represents the preferences of the decision maker and is assumed known. We allow that the payoff $U(a,y,x)$ is a nontrivial function of x. The table below illustrates the payoff function under $X = x$ with different combinations of (a,Y).

	State	
Action	$Y = 1$	$Y = -1$
$a = 1$	$U(1,1,x)$	$U(1,-1,x)$
$a = -1$	$U(-1,1,x)$	$U(-1,-1,x)$

As a primary application of this setting, we may regard a as a forecast of the outcome of a future random variable Y, or alternatively as an action taken based on the predicted binary outcome of Y. Then $U(a,y,x)$ is the payoff when the forecast or action is a, the realized value of Y is y, and the covariate vector is equal to x. In this application, we expect that

$$U(1,1,x) > U(1,-1,x) \text{ and } U(-1,-1,x) > U(-1,1,x) \text{ for all } x \in \mathscr{X}. \tag{2.3}$$

That is, a correct prediction delivers a higher payoff than an incorrect prediction.

As a second application, our setting can be cast as a 2×2 game where Nature plays Y and the decision maker plays a. More specifically, Nature plays a mixed strategy: for a given $X = x$, Nature plays $Y = 1$ with probability $P(x)$ and plays $Y = -1$ with probability $1 - P(x)$. In this case, (2.3) states that there is no dominating strategy for the decision maker.

We formalize (2.3) along with a self-explanatory technical condition as an assumption below.

Assumption 2.2.1 *(i) For all $x \in \mathscr{X}$,*

$$U(1,1,x) - U(-1,1,x) > 0$$

and

$$U(-1,-1,x) - U(1,-1,x) > 0;$$

(ii) for all $(a,y) \in \{-1,1\}^2$, $U(a,y,\cdot)$ is Borel measurable.

2.2.2 Utility Maximizing Actions

Given $X = x$, a decision maker's action is optimal if it maximizes her conditional expected utility, i.e., a^* is optimal if

$$a^* \in \arg\max_a E\left[U(a,Y,X)\,|X = x\right]. \tag{2.4}$$

Here a^* depends on the observed value x. To signify this, we write it as $a^*(x)$. We can alternatively formulate the decision maker's problem in terms of a loss function. We think about the loss of an action a as the amount by which the resulting utility differs from that of a perfect forecast if Y were known when the decision is made. Given Assumption 2.2.1(i), a perfect forecast would entail setting the category of action a to that of Y; we denote this unobtainable action based on the realization of Y by a_R. To motivate the form of the loss function, note that (2.4) is equivalent to

$$a^* \in \arg\min_a E\left[U(a_R,Y,X) - U(a,Y,X)|X = x\right]. \tag{2.5}$$

With $a_R = Y$ by Assumption 2.2.1(i), we define the loss function $\ell : \{-1,1\}^2 \times \mathscr{X} \to \mathbb{R}$ by

$$\ell(a,y,x) = U(y,y,x) - U(a,y,x). \tag{2.6}$$

This utility-induced loss function is called the *point-forecast/point-realization loss function* in Granger and Machina (2006). Clearly,

$$\ell(a,y,x) = \begin{cases} 0, & \text{if } a = y \\ U(y,y,x) - U(a,y,x) > 0, & \text{if } a \neq y. \end{cases}$$

In general, $\ell(a,y,x) \neq \ell(y,a,x)$, and so the loss function is not symmetric. In terms of the loss function, we have

$$a^* \in \arg\min_a E\left[\ell(a,Y,X)|X = x\right]. \tag{2.7}$$

We can now derive a solution of (2.7) (equation (2.9) below), which is also obtained in Elliott and Lieli (2013). When $X = x$ and $a = 1$, the expected loss is

$$E[\ell(1,Y,X)|X = x] = (1 - P(x))\ell(1,-1,x).$$

When $X = x$ and $a = -1$, the expected loss is

$$E[\ell(-1,Y,X)|X = x] = P(x)\ell(-1,1,x).$$

Now, if we let

$$\begin{aligned} b(x) &= \ell(1,-1,x) + \ell(-1,1,x), \\ c(x) &= \frac{\ell(1,-1,x)}{b(x)} = \frac{\ell(1,-1,x)}{\ell(1,-1,x) + \ell(-1,1,x)}, \end{aligned} \tag{2.8}$$

then a little algebra shows that an optimal decision rule, i.e., the one that obtains the lowest possible expected loss, is to set $a^*(x) = 1$ if and only if $P(x) > c(x)$. This can be written as

$$a^*(x) = \operatorname{sign}\left[P(x) - c(x)\right], \tag{2.9}$$

where $\operatorname{sign}(z) = 1$ for $z > 0$ and $\operatorname{sign}(z) = -1$ for $z \leq 0$.

For intuition, under Assumption 2.2.1 and provided that $P(x) < 1$, $a^*(x)$ in (2.9) can be restated as setting $a = 1$ if and only if

$$\frac{P(x)}{1 - P(x)} > \frac{\ell(1,-1,x)}{\ell(-1,1,x)} = \frac{U(-1,-1,x) - U(1,-1,x)}{U(1,1,x) - U(-1,1,x)}.$$

If we think of a as a prediction of Y based on $X = x$, then $\ell(1,-1,x) = U(-1,-1,x) - U(1,-1,x)$ is the *ex post* missed utility from a false positive prediction (i.e., take $a = 1$ when $Y = -1$) and $\ell(-1,1,x) = U(1,1,x) - U(-1,1,x)$ is the *ex post* missed utility from a false negative prediction (i.e., take $a = -1$ when $Y = 1$). The optimal decision rule sets $a = 1$ only when the odds ratio of the event $Y = 1$ relative to the event $Y = 0$ is greater than the false positive to false negative loss ratio. As the relative cost of a false positive gets larger, a greater odds ratio is required for an optimal utility-based decision rule to permit the action $a = 1$.

In terms of $b(x)$ and $c(x)$, the point-realization loss function in (2.6) can be written as

$$\ell(a,y,x) = \psi(x,y) \cdot 1\{y \neq a\}, \tag{2.10}$$

where

$$\psi(x,y) = b(x)\left[\frac{y+1}{2} - yc(x)\right] = U(y,y,x) - U(-y,y,x) > 0. \tag{2.11}$$

This can be easily verified. Therefore,

$$a^* \in \arg\min_a E\left[\psi(X,Y)1\{Y \neq a\}|X = x\right]. \tag{2.12}$$

The decision maker knows the payoff function $U(a,y,x)$ and hence $b(x), c(x)$, and $\psi(x,y)$. She does not know $P(x)$, the only piece of information that is still missing in solving the above minimization problem. To make an optimal decision, she has to estimate $P(x)$ based on the sample $\{(X_i,Y_i)\}_{i=1}^n$. One of her options would be to choose a proxy $m(x)$ for the unknown $P(x)$ from some class of functions. Her task is then to learn the most suitable m for a decision rule of the form $a(x) = \text{sign}[m(x) - c(x)]$. In considering such options, we will maintain the following additional sampling and distributional assumptions.

Assumption 2.2.2 *(i)* $\{(X_i,Y_i)\}_{i=1}^n$ *is an iid sample; (ii)* $X_i \in \mathscr{X}$ *and* $Y_i \in \{-1,1\}$; *(iii) The joint distribution function of* (X,Y) *is* $P(X,Y)$ *where* $P(X,Y)$ *is a probability measure over* $(\mathscr{X} \times \{-1,1\}, \mathscr{B}_x \otimes \mathscr{B}_y)$ *where* $\mathscr{B}_x$ *is the Borel* σ*-algebra associated with* $\mathscr{X}$ *and* $\mathscr{B}_y$ *consists of all subsets of* $\{-1,1\}$; *(iv) There exists some* $K_\psi > 0$ *such that*

$$E \exp\left\{\lambda^2 \psi(X,Y)^2\right\} \leq \exp\left\{K_\psi^2 \lambda^2\right\} \text{ for all } \lambda \text{ such that } |\lambda| \leq \frac{1}{K_\psi}.$$

The condition on the moment generating function in Assumption 2.2.2(iv) specifies that the random variable $\psi(X,Y)$ is sub-Gaussian (c.f. Proposition 2.5.2 (iii) and Definition 2.5.6 of Vershynin (2018)). Given that $\psi(x,y) = U(y,y,x) - U(-y,y,x)$, this assumption requires that the payoffs from a correct decision (or alternatively, the costs from an incorrect decision) must be sub-Gaussian. This is a fairly mild requirement and accommodates, for example, any underlying utility function that is bounded, a condition that is assumed in Elliott and Lieli (2013) and Su (2020). Here benefits and costs of correct or incorrect decisions do not have to be bounded provided that the tails of the distribution decay exponentially.

In terms of the resulting action rule $a_{m^*}(x) = \text{sign}[m^*(x) - c(x)]$, the conditional optimization problem (2.12) is equivalent to the unconditional optimization problem

$$m^* \in \underset{m \in \mathscr{M}}{\arg\min}\, E\left\{\psi(X,Y)\, 1\{Y \neq \text{sign}[m(X) - c(X)]\}\right\}, \tag{2.13}$$

where $\mathscr{M}$ is the space of all measurable functions from $\mathscr{X}$ to $\mathbb{R}$. To implement the optimal m^*, the decision maker could solve the sample version of the above problem,

$$\hat{m}^* \in \arg\min_{m\in\mathscr{M}} \frac{1}{n}\sum_{i=1}^{n} \psi\left(X_i, Y_i\right) 1\left\{Y_i \neq \operatorname{sign}\left[m(X_i) - c(X_i)\right]\right\},$$

and let

$$a_{\hat{m}^*}\left(x\right) = \operatorname{sign}\left[\hat{m}^*(x) - c(x)\right].$$

The M estimator $\hat{m}^*$ is motivated from utility maximization, and we will refer to it as the maximum utility (MU) estimator. The MU estimator is clearly different from the maximum likelihood estimator defined as

$$\hat{m}^*_{MLE} = \arg\max_{m\in\mathscr{M}} \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{Y_i+1}{2}\log m\left(X_i\right) + \left(1 - \frac{Y_i+1}{2}\right)\log\left[1 - m(X_i)\right]\right\},$$

where we have assumed that $m(X_i) \in \left(0,1\right)$.[1] The likelihood function is motivated statistically without accounting for the payoff differences under different actions and states of the world.

Implementation of the optimal strategy requires searching over the whole space of measurable functions $\mathscr{M}$. This is a formidable task. In addition, such a method may not generalize well. In practice, we restrict attention to a parameterized subclass of $\mathscr{M}$. Denote this collection of models by $\mathscr{M}_\Theta \subset \mathscr{M}$ where each model $m(x,\theta) \in \mathscr{M}_\Theta$ is determined by parameters $\theta \in \Theta$ and $\Theta \subset \mathbb{R}^q$ is the parameter space with potentially $q \neq d$, where d is the dimension of $\mathscr{X}$. We delay specifying the functional form of $m(x,\theta)$ for now. The MU estimator over $\mathscr{M}_\Theta$ selects the model parameter $\hat{\theta}$ by solving

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \psi\left(X_i, Y_i\right) 1\left\{Y_i \neq \operatorname{sign}[m(X_i,\theta) - c(X_i)]\right\}.$$

Such an estimator has been considered in Elliott and Lieli (2013). In the special case that the loss

[1] If this is not the case, we can take a transform such as the logistic transform so that the transformed version is in $(0,1)$.

functions $\ell(1,-1,x)$ and $\ell(-1,1,x)$ are equal to the same constant function, we have $c(x) = 1/2$ and $\psi(x,y) = [\ell(1,-1,x)+\ell(-1,1,x)]/2$, which is also a constant function. Hence,

$$\hat{\theta} \in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} 1\left\{Y_i \neq \operatorname{sign}[m(X_i,\theta) - c(X_i)]\right\}.$$

In this case, the MU estimator reduces to the maximum score estimator of Manski (1975, 1985). Su (2020) considers model selection in the MU framework. There, model selection is based on a penalized MU estimator where the additive penalty regularizes the complexity of the model class and controls the generalization error.

A key observation from Elliott and Lieli (2013) is that m^* and $\hat{m}^*$ may not be unique. Consider the sample problem as an example. If $\hat{m}^*$ is a solution, then any function $\hat{m}(x)$ that satisfies

$$\operatorname{sign}[\hat{m}^*(x) - c(x)] = \operatorname{sign}[\hat{m}(x) - c(x)]$$

is also a solution. Each crossing point of $P(x)$ and $c(x)$ corresponds to a region of $\mathscr{X}$ where $\hat{m}^*$ and $\hat{m}$ may disagree out of sample even if both achieve the same in-sample empirical utility. This provides an incentive to consider ensemble methods. In the presence of multiple solutions, it is reasonable to average or aggregate models with high empirical utility rather than trying to select a solution.

2.2.3 PAC-Bayesian Framework

Instead of model selection, we consider model aggregation in this paper. We do so within the PAC-Bayesian framework. In this subsection, we introduce some definitions and concepts central to this approach before considering statistical properties of the resulting decision rules in Section 2.3.

Most generally, we work with $\mathscr{R}_\Theta$, a parameterized subclass of the set of measurable functions from $\mathscr{X}$ to $\{-1,1\}$, characterized by a parameter space Θ. The typical example we deal with here and in our simulations is the setting where

$$\mathscr{R}_\Theta = \{\text{sign}\,[m(x,\theta) - c(x)] : m \in \mathscr{M}_\Theta\}, \tag{2.14}$$

where again $\mathscr{M}_\Theta$ is a parameterized subclass of the space of measurable functions $m : \mathscr{X} \to \mathbb{R}$ that are characterized by the parameter space $\Theta \subset \mathbb{R}^q$ associated with q model parameters.

For actions $a\,(x,\theta) \in \mathscr{R}_\Theta$ (determined by $\theta \in \Theta$), with some abuse of notation, we denote the utility-induced, point-realization loss by

$$\ell(\theta,y,x) = \psi(x,y)1\,\{y \neq a\,(x,\theta)\},$$

where $\psi(x,y)$ is defined in (2.11). Additionally, for any $\theta \in \Theta$, define the risk function $R(\theta)$ and its empirical counterpart $R_n(\theta)$ by

$$R(\theta) = E[\ell(\theta,Y,X)], \tag{2.15}$$

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta,Y_i,X_i). \tag{2.16}$$

Whereas the MU approach selects a single $\hat{\theta} \in \Theta$ by minimizing $R_n(\theta)$ over Θ, here we will place a non-negative weighting on each θ in the form of a probability measure on Θ and then take actions based on aggregation over all possible models. The goal is to construct a probability measure $\rho(\cdot)$ on Θ that may depend on the sample $\{(X_i,Y_i)\}_{i=1}^n$. The PAC-Bayesian framework allows us to identify bounds on functionals of $R(\theta)$ that depend on $R_n(\theta)$ and hold with high probability. These bounds can then guide the choice of ρ. We will need to integrate over both the sample space and the parameter space, and we make the following assumption.

Assumption 2.2.3 *(i) $(\Theta, \mathscr{B}_\theta)$ is a measurable space where $\mathscr{B}_\theta$ is the standard σ-algebra on Θ and is countably generated; (ii) $(\theta,x) \mapsto a\,(x,\theta) : (\Theta \times \mathscr{X}, \mathscr{B}_\theta \otimes \mathscr{B}_x) \to (\{-1,1\}, \mathscr{B}_a)$ is a measurable function where $\mathscr{B}_a = \mathscr{B}_y$.*

Assumption 2.2.3 contains some technical conditions that address measurability concerns. By a probability measure $\rho(\cdot)$ on Θ that may be sample dependent, we mean a regular conditional probability measure $\rho(z,\cdot)$ where $z \in (\mathscr{X} \times \{-1,1\})^{\times n}$. That is, for any fixed $S \in \mathscr{B}_\theta$, $\rho(z,S) : ((\mathscr{X} \times \{-1,1\})^{\times n}, (\mathscr{B}_x \otimes \mathscr{B}_y)^{\otimes n}) \to \mathbb{R}_+$ is measurable in z and for any fixed z, the map $S \mapsto \rho(z,S) : \mathscr{B}_\theta \to \mathbb{R}_+$ is a probability measure. For conciseness, we suppress the potential reliance of ρ on the particular sample set z. Given some deterministic probability measure π, we will work with the Kullback-Leibler (KL) divergence between π and ρ,

$$D_{\mathrm{KL}}(\rho,\pi) = \begin{cases} \int_\Theta \log\left[\frac{d\rho}{d\pi}(\theta)\right] d\rho(\theta), & \text{if } \rho \ll \pi \\ \infty, & \text{else.} \end{cases}$$

We will consider only the case that $\rho \ll \pi$ (a.s.) in this paper. The requirement in Assumption 2.2.3 that $\mathscr{B}_\theta$ is countably generated serves to ensure that objects such as $D_{\mathrm{KL}}(\rho,\pi)$ are measurable. For further measure-theoretic consideration, we refer the reader to Catoni (2004), in particular Proposition 1.7.1 and its proof on pages 50-54. There the measurability of $D_{\mathrm{KL}}(\rho,\pi)$ when ρ and π may be regular conditional probability measures is demonstrated under conditions that are met by our assumptions.

Given a probability measure $\rho(\cdot)$ over Θ, there are a few ways to form a decision rule. Among them, the Gibbs method and the majority vote method are widely used. The Gibbs method associated with ρ draws a value, say $\theta_\circ$, randomly according to ρ and then takes the action based on $\theta_\circ$. Mathematically, we let $\theta_\circ \sim \rho$ and we take

$$a_{G,\rho}(x) = a_{\theta_\circ}(x).$$

That is, we play a mixed strategy based on the distribution ρ. With some abuse of the notation, the average risk of the Gibbs method associated with ρ is

$$R\left(a_{G,\rho}\right) = \int_{\Theta} R(\theta) d\rho(\theta) = E_{\theta\sim\rho} E_{X,Y\sim P(X,Y)} \psi(X,Y) 1\left\{Y \neq a(X,\theta)\right\}, \tag{2.17}$$

which is referred to as the Gibbs risk in the literature. Above, $E_{\theta\sim\rho}$ is the expectation with respect to the distribution of θ, and $E_{X,Y\sim P(X,Y)}$ is the expectation with respect to the distribution of (X,Y). We adopt the same convention hereafter.

The Gibbs risk $R\left(a_{G,\rho}\right)$ is the expectation of the risk function $R(\theta)$ under measure $\rho(\cdot)$. It is thus a linear functional of $\rho(\cdot)$. More precisely, if $\rho = \alpha\rho_1 + (1-\alpha)\rho_2$ for some ρ_1 and ρ_2 and a constant α, then $R\left(a_{G,\rho}\right) = \alpha R\left(a_{G,\rho_1}\right) + (1-\alpha) R\left(a_{G,\rho_2}\right)$. The linearity makes the Gibbs risk more amenable to theoretical analysis.

For the majority vote method, which is also called the Bayes method, the action is defined according to

$$a_{B,\rho}(x) = \operatorname{sign}\left\{E_{\theta\sim\rho} a(x,\theta)\right\}.$$

Such a method aggregates the actions $\{a(x,\theta) : \theta \in \Theta\}$ to obtain the prevailing action. For intuition, suppose that $\theta_1, \ldots, \theta_N$ are N i.i.d. draws from ρ and consider the action

$$a_{B,N}(x) = \operatorname{sign}\left\{\frac{1}{N}\sum_{j=1}^{N} a(x,\theta_j)\right\}.$$

Provided that $E_{\theta\sim\rho} a(x,\theta) \neq 0$, we have $a_{B,N}(x) \overset{\text{a.s.}}{\to} a_{B,\rho}(x)$ as $N \to \infty$ for each x. Note that $a_{B,N}(x) = 1$ if and only if more than half of the actions $\{a(x,\theta_j)\}_{j=1}^{N}$ are equal to 1 so this is akin to a weighted majority vote of the parameter values in Θ. The risk of the majority vote (also called the Bayes risk) associated with ρ is defined by

$$\begin{aligned} R\left(a_{B,\rho}\right) &= E_{X,Y\sim P(X,Y)} \psi(X,Y) 1\left\{Y \neq a_{B,\rho}(X)\right\} \\ &= E_{X,Y\sim P(X,Y)} \psi(X,Y) 1\left\{Y \neq \operatorname{sign}\left\{E_{\theta\sim\rho} a(X,\theta)\right\}\right\}. \end{aligned}$$

The Bayes risk is clearly not linear in ρ.

In practice, the majority vote method or the Bayes method delivers numerically more stable results than the Gibbs method, but the latter is easier to analyze. However, the Bayes risk is upper bounded by twice the Gibbs risk as shown in the following lemma.

Lemma 2.2.1 *Let Assumption 2.2.1, Assumptions 2.2.2(ii) and (iii), and Assumption 2.2.3 hold. Then, for any probability measure ρ on Θ,*

$$R\left(a_{B,\rho}\right) \leq 2R\left(a_{G,\rho}\right).$$

Lemma 2.2.1 extends the "factor 2" bound for the majority vote method in the machine learning literature to the utility-based, point-realization loss setting. This property is well documented in the case of $0/1$ loss (e.g., Langford and Shawe-Taylor (2003), McAllester (2003a), and Germain et al. (2015)). Here, we use Lemma 2.2.1 only to justify using the Gibbs risk as a surrogate for the majority vote risk. The loose bound in the lemma is enough for this purpose. Langford and Shawe-Taylor (2003) show that the factor of 2 can sometimes be reduced to $(1+\varepsilon)$ for some small $\varepsilon > 0$. Lacasse et al. (2006) and Germain et al. (2015) show that tighter bounds on $R(a_{B,\rho})$ can be obtained in the $0/1$ loss setting and in a related loss variant.

To choose ρ to guide our decisions, we follow the PAC-Bayesian approach. Let $\mathscr{P}(\Theta)$ be the set of probability measures on $(\Theta, \mathscr{B}_\theta)$. The first ingredient is a reference or prior probability measure π. We make the following assumption:

Assumption 2.2.4 $\pi \in \mathscr{P}(\Theta)$ *is a (deterministic) probability measure that does not depend on the sample.*

We will denote the set of probability measures on $(\Theta, \mathscr{B}_\theta)$ that are absolutely continuous with respect to π by $\mathscr{P}_\pi(\Theta)$. Assumption 2.2.4 is essential in PAC-Bayesian analysis. For example, our analysis will involve the sample version of the Gibbs risk, defined for $\rho \in \mathscr{P}(\Theta)$

by

$$R_n(a_{G,\rho}) = \int_\Theta R_n(\theta) d\rho = \sum_{i=1}^n \int_\Theta \ell(\theta, Y_i, X_i) d\rho(\theta). \tag{2.18}$$

If ρ is derived from $\{X_i, Y_i\}_{i=1}^n$, (2.18) is difficult to work with because $\int_\Theta \ell(\theta, Y_i, X_i) d\rho(\theta)$ is not iid and so $R_n(a_{G,\rho})$ is not a sum of iid terms. However, for any measurable function $A(\theta)$, the so-called change-of-measure inequality states that for *any* $\rho \in \mathscr{P}_\pi(\Theta)$,

$$\int_\Theta A(\theta) d\rho(\theta) \leq \log\left[\int_\Theta \exp\left(A(\theta)\right) d\pi(\theta)\right] + D_{\text{KL}}(\rho, \pi), \tag{2.19}$$

provided the integrals are well defined. When both $\rho(\cdot)$ and $A(\cdot)$ depend on the sample and exhibit complicated dependence, it may not be easy to control $\int_\Theta A(\theta) d\rho(\theta)$. But when π does not depend on the sample, (2.19) can provide a manageable upper bound. Although the change of measure inequality is simple and easy to prove, it is foundational to the PAC-Bayesian approach. See McAllester (2003b) and references therein for further discussion. (2.19) is stated below as Corollary 2.2.1(b), and a proof is given in the appendix. Some choices for π are discussed in Sections 2.3.2 and 2.4.

Given the pre-specified π, we choose ρ to minimize the sample Gibbs risk $R_n(a_{G,\rho})$ in (2.18), subject to the constraint that ρ is not too different from π. We utilize the KL divergence to measure the difference between two probability measures. Mathematically, we solve the constrained minimization problem:

$$\min_{\rho \in \mathscr{P}_\pi(\Theta)} \int_\Theta R_n(\theta) d\rho(\theta) \;\; \text{s.t. } D_{\text{KL}}(\rho, \pi) \leq C,$$

for some constant C. Alternatively, we use the Lagrangian form and solve the unconstrained minimization problem

$$\min_{\rho \in \mathscr{P}_\pi(\Theta)} \left[\int_\Theta R_n(\theta) d\rho(\theta) + \frac{1}{\lambda} D_{\text{KL}}(\rho, \pi)\right], \tag{2.20}$$

where $\lambda > 0$ is a constant. Theoretical justification for this choice of optimization problem is given in Section 2.3.

Let $\mathscr{M}(\Theta)$ be the set of measurable functions on $(\Theta, \mathscr{B}_\theta)$ and

$$\mathscr{M}_b^\pi(\Theta) = \left\{ A : A \in \mathscr{M}(\Theta) \text{ and } \int_\Theta \exp(A(\theta))\, d\pi(\theta) < \infty \right\},$$

which is a subset of $\mathscr{M}(\Theta)$ that has a finite exponential moment under π. To obtain a closed-form solution to 2.20, we provide the following lemma and corollary, which will also be used repeatedly for establishing other results.

Lemma 2.2.2 *For $\pi \in \mathscr{P}(\Theta)$ and $A \in \mathscr{M}(\Theta)$ such that $-A \in \mathscr{M}_b^\pi(\Theta)$, let $\rho_{A,\pi} \in \mathscr{P}_\pi(\Theta)$ be the probability measure on Θ with the Radon–Nikodym (RN) derivative with respect to π given by*

$$\frac{d\rho_{A,\pi}(\theta)}{d\pi(\theta)} = \frac{\exp(-A(\theta))}{\int_\Theta \exp\left(-A\left(\tilde{\theta}\right)\right) d\pi\left(\tilde{\theta}\right)}.$$

Then for any probability measure $\rho \in \mathscr{P}_\pi(\Theta)$ we have

$$\log\left[\int_\Theta \exp(-A(\theta))\, d\pi(\theta)\right] = -\left[\int_\Theta A(\theta)\, d\rho(\theta) + D_{\mathrm{KL}}(\rho, \pi)\right] + D_{\mathrm{KL}}\left(\rho, \rho_{A,\pi}\right). \quad (2.21)$$

Corollary 2.2.1 *(a) For A, π, ρ, and $\rho_{A,\pi}$ as in Lemma 2.2.2, we have*

$$\rho_{A,\pi} = \arg\min_{\rho \in \mathscr{P}_\pi(\Theta)} \left[\int_\Theta A(\theta)\, d\rho(\theta) + D_{\mathrm{KL}}(\rho, \pi)\right] \quad (2.22)$$

and

$$\min_{\rho \in \mathscr{P}_\pi(\Theta)} \left[\int_\Theta A(\theta)\, d\rho(\theta) + D_{\mathrm{KL}}(\rho, \pi)\right] = -\log\left[\int_\Theta \exp(-A(\theta))\, d\pi(\theta)\right].$$

(b) For any $\mathscr{A}(\cdot) \in \mathscr{M}_b^\pi(\Theta)$, $\pi \in \mathscr{P}(\Theta)$, $\rho \in \mathscr{P}_\pi(\Theta)$,

$$\int_\Theta \mathscr{A}(\theta)\, d\rho(\theta) \leq \log\left[\int_\Theta \exp(\mathscr{A}(\theta))\, d\pi(\theta)\right] + D_{\mathrm{KL}}(\rho, \pi).$$

Lemma 2.2.2 and Corollary 2.2.1(a) provide a closed-form solution to the minimization problem in (2.20). Let

$$\hat{\rho}_\lambda := \arg \min_{\rho \in \mathscr{P}_\pi(\Theta)} \left[\int_\Theta R_n(\theta)\, d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi) \right]. \tag{2.23}$$

Then it follows from Corollary 2.2.1(a) that $\hat{\rho}_\lambda = \rho_{\lambda R_n, \pi}$. Given λ and π, $\hat{\rho}_\lambda$ will be our primary choice of probability measure for deriving decision rules through a majority vote or Gibbs method. We present this as a definition.

Definition 2.2.1 *$\hat{\rho}_\lambda$ is a (random) probability measure on Θ with the following RN derivative with respect to π :*

$$\frac{d\hat{\rho}_\lambda}{d\pi}(\theta) = \frac{\exp\left[-\lambda R_n(\theta)\right]}{\int_\Theta \exp\left[-\lambda R_n\left(\tilde{\theta}\right)\right] d\pi\left(\tilde{\theta}\right)}.$$

$\hat{\rho}_\lambda$ is sometimes called the Gibbs posterior. From a Bayesian perspective, we may regard π as the prior distribution for the parameter $\theta \in \Theta$ and $\hat{\rho}_\lambda$ as the posterior distribution. Such a Bayesian interpretation may help us understand the approach, but it is not necessary. In fact, this interpretation is valid only if $\exp\left[-\lambda R_n(\theta)\right]$ is proportional to a likelihood function. The approach we use is a frequentist one, and $\exp\left[-\lambda R_n(\theta)\right]$ does not have to be a likelihood function. The definition of $\hat{\rho}_\lambda$ is motivated from the minimization problem in (2.23), not from any Bayesian principle. In particular, there does not have to be a likelihood function or a complete model. All we need is the empirical risk based on the utility-based loss function. Also, π does not have to be a prior distribution. It can be any distribution that does not depend on the sample. However, for easy references, we may still refer to π as the prior and $\hat{\rho}_\lambda$ as the posterior. More generally, any ρ determined from the sample may be referred to as a posterior distribution.

The probability measure $\hat{\rho}_\lambda$ can be regarded as an adjusted version of π. Consider two parameters $\theta_1 \in \Theta$ and $\theta_2 \in \Theta$. If $R_n(\theta_1) < R_n(\theta_2)$, then $\exp\left[-\lambda R_n(\theta_1)\right] > \exp\left[-\lambda R_n(\theta_2)\right]$ for any $\lambda > 0$. Hence, relative to $\pi, \hat{\rho}_\lambda$ assigns more weights to θ_1 than to θ_2. The distributional adjustment, therefore, favors the parameter value that delivers a smaller in-sample empirical risk. The degree of adjustment is determined by the tuning parameter λ. On the one hand, if λ approaches zero, then $\hat{\rho}_\lambda$ approaches π, and there will be no adjustment. On the other hand, if $\lambda \to +\infty$, then $\hat{\rho}_\lambda$ assigns all weights to the minimizers of $R_n(\theta)$, provided that the minimizers are in the support of the prior π. We will investigate the choice of λ in subsequent sections.

2.3 PAC-Bayesian Analysis Under Utility-Based Loss

In this section, we derive PAC-Bayesian bounds on the Gibbs risk and oracle inequalities for decision rules based on $\hat{\rho}_\lambda$ in Definition 2.2.1 for the utility-induced loss setting. The bounds provide justification for focusing on the minimization problem in (2.20). They are non-asymptotic training set bounds that hold for a user-specified confidence level. The oracle inequalities illustrate a sense in which $\hat{\rho}_\lambda$ is close to the probability measure we would select if $R(\theta)$ were known. We also consider a constrained version of the problem in (2.20) which illustrates the mechanics of the methodology and produces decision rules with similarities to support vector machines. Lastly, we consider the formulation when one is interested in aggregating multiple decision model classes.

For a probability measure ρ on Θ that may depend on the sample, an integral step in PAC-Bayesian analysis is to establish an upper bound for $D[R(a_{G,\rho}), R_n(a_{G,\rho})]$ where $D : \mathbb{R}_+^2 \to \mathbb{R}$ is a measure of the difference between the Gibbs risk $R(a_{G,\rho})$ defined in (2.17) and its empirical counterpart $R_n(a_{G,\rho})$ defined in (2.18). We will often focus on the case $D(r_1, r_2) = r_1 - r_2$, i.e., when

$$D\left[R\left(a_{G,\rho}\right), R_n\left(a_{G,\rho}\right)\right] = \int_\Theta R(\theta)\, d\rho(\theta) - \int_\Theta R_n(\theta)\, d\rho(\theta).$$

Let $\varepsilon > 0$ be a small constant. The initial aim is to establish the following result: for some upper bound $B_n(\pi, \rho, \varepsilon)$ we have

$$\Pr\left\{D\left[R\left(a_{G,\rho}\right), R_n\left(a_{G,\rho}\right)\right] \le B_n(\pi,\rho,\varepsilon) \text{ for all } \rho \in \mathscr{P}_\pi(\Theta) \text{ simultaneously}\right\} \ge 1 - \varepsilon. \quad (2.24)$$

We can use such a bound to choose $\hat{\rho} \in \mathscr{P}_\pi(\Theta)$ so that the Gibbs risk of $\hat{\rho}$, $R(a_{G,\hat{\rho}})$, is minimized with high probability. We will see that this leads to the minimization problem in (2.20).

For a given $D(\cdot,\cdot)$, we can regard $D[R(a_{G,\hat{\rho}}), R_n(a_{G,\hat{\rho}})]$ as a measure of the generalization error under the Gibbs method for $\hat{\rho}$. If $B_n(\pi,\hat{\rho},\varepsilon)$ decays to zero for any $\varepsilon > 0$ as n increases, then the above inequality implies a low generalization error with high probability (i.e., with probability at least $1-\varepsilon$ for any small ε). In this case, we say that $R_n(a_{G,\hat{\rho}}) = \int_\Theta R_n(\theta)d\hat{\rho}(\theta)$ is probably (the high probability part) and approximately correct (the low generalization error part) for $R(a_{G,\hat{\rho}}) = \int_\Theta R(\theta)d\hat{\rho}(\theta)$. The PAC framework, introduced by Valiant (1984), evaluates learning mechanisms via the probability (prescribing a confidence level) that the resulting rule will approximate an optimal rule at some level of accuracy. As noted in Shalev-Shwartz and Ben-David (2014), which includes an excellent introduction to PAC analysis, this framework has broad appeal, has been extended in scope (e.g. Haussler (1992)), and has been utilized in several foundational analyses (e.g. Vapnik (1982), Vapnik (1992), and Vapnik (2013)). In the PAC-Bayesian framework, rather than centering attention on learning mechanisms that settle on a particular instance in the parameter space, the focus rests on PAC statements for objects concerning distributions over models or model parameters. The approach then has flavors of both Probably Approximately Correct (PAC) learning and Bayesian learning. Hence it can be called PAC-Bayesian learning. As we discussed previously, the Bayesian part is a misnomer, and we use “PAC-Bayesian” in the absence of a better term.

2.3.1 Bounds and Oracle Inequalities for the Decision Rule

Here we establish PAC-Bayesian and oracle bounds under Assumptions 2.2.1 – 2.2.4. We begin with the following bound of the form in (2.24).

Theorem 2.3.1 *Let Assumptions 2.2.1, 2.2.2, and 2.2.3 hold. Let* $D : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ *be convex over the range of* $(\psi(x,y), \psi(x,y))$ *where* ψ *is defined in* (2.11) *and depends on the utility function* $U(a,y,x)$*. Assume there exists a function* $f(\lambda,n)$ *and an interval* $I \subseteq \mathbb{R}_+^* = \{\lambda \in \mathbb{R} :$

$\lambda > 0\}$ *such that for all* $\lambda \in I$,

$$\int_{\Theta} E \exp\left(\lambda D\left[R(\theta), R_n(\theta)\right]\right) d\pi(\theta) \leq \exp(f(\lambda, n)). \tag{2.25}$$

Then for any $\varepsilon \in (0,1]$,

$$\Pr\left\{D\left[R\left(a_{G,\rho}\right), R_n\left(a_{G,\rho}\right)\right] \leq \frac{f(\lambda, n) + \log\frac{1}{\varepsilon} + D_{\mathrm{KL}}(\rho, \pi)}{\lambda} \textit{ for all } \rho \in \mathscr{P}_{\pi}(\Theta) \textit{ simultaneously}\right\}$$
$$\geq 1 - \varepsilon. \tag{2.26}$$

There is a fairly well established path to results like Theorem2.3.1 in the literature. For example, Bégin et al. (2016) lays out a blueprint for deriving such bounds in the $0/1$ loss setting that is general enough to encompass many results identified in the previous literature. The above bound combines elements of Theorem 4.2 in Alquier et al. (2016) and Theorem 18 in Germain et al. (2015). Theorem 2.3.1 is proved in the Appendix. Alquier et al. (2016) refer to condition (2.25) as the Hoeffding assumption. In situations where $D[R(\theta), R_n(\theta)]$ may become unbounded almost surely for certain values of θ, such a condition can allow for valid and nontrivial PAC-Bayesian bounds provided that $\pi(\theta)$ is chosen judiciously. We will also note that D in Theorem 2.3.2 may depend on λ provided that for each $\lambda \in I$ it is convex over the range of $(\psi(x,y), \psi(x,y))$. In our analysis, when D depends on λ in this way, it will be the case that the resulting high probability inequality simplifies so that the left-hand-side contains an object of interest and does not depend on λ.

To produce the main bounds and oracle inequalities of interest, we combine the above theorem with the following lemma.

Lemma 2.3.1 *Let Assumptions 2.2.1 – 2.2.4 hold.*

(a) For $s \in \{-1, 1\}$, *let* $D(r_1, r_2) = s(r_1 - r_2)$, *so that*

$$D\left[R(\theta), R_n(\theta)\right] = s\left(R(\theta) - R_n(\theta)\right).$$

Then for $\lambda > 0$, (2.25) *holds with*

$$f(\lambda,n) = \frac{\lambda^2 \left[K_\psi^2 + \mu_\psi^2\right]}{n},$$

where K_ψ is the constant in Assumption 2.2.2 and $\mu_\psi = E\psi(X,Y)$. Additionally, if

$$U_{\max} = \sup_{a,y,x} |U(a,y,x)| < \infty, \tag{2.27}$$

then for $\lambda > 0$, (2.25) *holds with*

$$f(\lambda,n) = \frac{\lambda^2 U_{\max}^2}{2n}.$$

(b) Assume (2.27) *holds. Let*

$$D(r_1,r_2) = \mathscr{F}(r_1) - r_2,$$

where

$$\mathscr{F}(r) := \mathscr{F}_{n,\lambda}(r) = -\frac{n}{\lambda}\log\left\{1 - \frac{r}{2U_{\max}}\left[1 - \exp\left(-\frac{2U_{\max}\lambda}{n}\right)\right]\right\}. \tag{2.28}$$

Then, for $\lambda > 0$, (2.25) *holds with*

$$f(\lambda,n) = 0.$$

(c) Assume (2.27) *holds. Let*

$$D(r_1,r_2) = \max\left\{r_1 - \frac{\lambda U_{\max}^2}{2n} - r_2,\ \mathscr{F}(r_1) - r_2\right\},$$

where $\mathscr{F}$ is defined as in (2.28)*. Then, for $\lambda > 0$,* (2.25) *holds with*

$$f(\lambda,n) = 0.$$

Theorem 2.3.1 combined with Lemma 2.3.1 produces the following result.

Theorem 2.3.2 *Under Assumptions 2.2.1 – 2.2.4, for $\lambda > 0$ and $\varepsilon \in (0,1]$ we have the following* *properties.*

(a) For $s \in \{-1, 1\}$, the following event occurs with probability at least $1-\varepsilon$ for all $\rho \in \mathscr{P}_\pi(\Theta)$ simultaneously:

$$\int_\Theta s\left[R(\theta) - R_n(\theta)\right] d\rho(\theta) \leq \frac{1}{\lambda}\left[\frac{\lambda^2}{n}\left(K_\psi^2 + \mu_\psi^2\right) + D_{\mathrm{KL}}(\rho, \pi) + \log\frac{1}{\varepsilon}\right]$$

where K_ψ is the constant in Assumption 2.2.2 and $\mu_\psi = E\psi(X,Y)$. If (2.27) *holds, then the term $(K_\psi^2 + \mu_\psi^2)$ can be replaced by $U_{\max}^2/2$.*

(b) If (2.27) *holds, then the following event occurs with probability at least $1-\varepsilon$ for all $\rho \in \mathscr{P}_\pi(\Theta)$ simultaneously:*

$$\int_\Theta R(\theta)\, d\rho(\theta) \leq \mathscr{F}_{n,\lambda}^{-1}\left(\int_\Theta R_n(\theta)\, d\rho(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi) + \frac{1}{\lambda}\log\frac{1}{\varepsilon}\right).$$

where $\mathscr{F}_{n,\lambda}^{-1}(r)$ is the inverse function of $\mathscr{F}_{n,\lambda}(r)$:

$$\mathscr{F}_{n,\lambda}^{-1}(r) = 2U_{\max}\frac{1-\exp\left(-\frac{\lambda}{n}\cdot r\right)}{1-\exp\left(-\frac{\lambda}{n}\cdot 2U_{\max}\right)}.$$

(c) Define

$$U_{\lambda,\pi,\rho}(\varepsilon) = \int_\Theta R_n(\theta) d\rho(\theta) + \frac{1}{\lambda}\left[\frac{\lambda^2 U_{\max}^2}{2n} + D_{\mathrm{KL}}(\rho,\pi) + \log\frac{1}{\varepsilon}\right], \tag{2.29}$$

$$U_{\lambda,\pi,\rho}^{\mathscr{F}}(\varepsilon) = \mathscr{F}_{n,\lambda}^{-1}\left(\int_\Theta R_n(\theta)\, d\rho(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi) + \frac{1}{\lambda}\log\frac{1}{\varepsilon}\right). \tag{2.30}$$

If (2.27) *holds, the following event occurs with probability at least $1-\varepsilon$ for all $\rho \in \mathscr{P}_\pi(\Theta)$ simultaneously:*

$$\int_\Theta R(\theta)\, d\rho(\theta) \leq \min\left\{U_{\lambda,\pi,\rho}(\varepsilon), U_{\lambda,\pi,\rho}^{\mathscr{F}}(\varepsilon)\right\}.$$

When $U_{\max} < \infty$, the bounds in Theorem 2.3.2(a), (b), and (c) can be computed from the sample for a learned ρ, be it of the form in Definition 2.2.1 or that in Section 2.3.2 or some other form. In the $U_{\max} < \infty$ setting, part (c) can provide an improvement over the bounds in part (a) and (b) which are, respectively, similar in form to bounds in Alquier et al. (2016) and Catoni (2007). Setting $s = 1$ in Theorem 2.3.2(a), we obtain, with probability at least $1 - \varepsilon$ for all $\rho \in \mathscr{P}_\pi(\Theta)$ simultaneously:

$$\int_\Theta R(\theta)\, d\rho(\theta) \leq \left[\int_\Theta R_n(\theta)\, d\rho(\theta) + \frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi)\right] + \frac{1}{\lambda}\left[\frac{\lambda^2}{n}\left(K_\psi^2 + \mu_\psi^2\right) + \log\frac{1}{\varepsilon}.\right]$$

The above bound and the bound in 2.3.2(b) are slight variants of one another. Note that for a given λ, if we choose ρ to minimize the upper bound for $R(a_{G,\rho}) = \int_\Theta R(\theta) d\rho(\theta)$ in either of the inequalities we are led back to the minimization problem in (2.20). The bound in Theorem 2.3.2(b) is similar in form to Theorem 1.2.6 in Catoni (2007) for the $0/1$ loss. It is recovered from the distance measure D in Lemma 2.3.1(b) similarly to Germain et al. (2009) who focus on the $0/1$ loss setting.

When $s = 1$, Theorem 2.3.2(a) gives us

$$\Pr\left\{\int_\Theta [R(\theta) - R_n(\theta)]\, d\rho(\theta) \leq B_n(\pi, \rho, \varepsilon) \text{ for all } \rho \in \mathscr{P}_\pi(\Theta) \text{ simultaneously}\right\} \geq 1 - \varepsilon,$$

for

$$B_{n,\lambda}(\pi, \rho, \varepsilon) = \frac{\lambda}{n}\left(K_\psi^2 + \mu_\psi^2\right) + \frac{1}{\lambda}\left[\log\frac{1}{\varepsilon} + D_{\mathrm{KL}}(\rho, \pi)\right].$$

Setting λ proportional to $n^{1/2}$ yields the following best rate of the PAC bound $B_{n,\lambda}(\pi, \rho, \varepsilon)$:

$$B_{n,\lambda}(\pi, \rho, \varepsilon) = O_p\left(\frac{1}{\sqrt{n}}\right).$$

On the other hand, for the function $\mathscr{F}_{n,\lambda}^{-1}(\cdot)$ in Theorem 2.3.2(b), we have, using $\exp(x) \geq 1 + x$ for all $x \in \mathbb{R}$,

$$\mathscr{F}_{n,\lambda}^{-1}(r) = 2U_{\max}\frac{1 - \exp\left(-\frac{\lambda}{n}\cdot r\right)}{1 - \exp\left(-\frac{\lambda}{n}\cdot 2U_{\max}\right)} \leq \frac{C_n}{1 - \exp(-C_n)} r$$

where $C_n = \frac{\lambda}{n} \cdot 2U_{\max}$. Hence, Theorem 2.3.2(b) implies that

$$\Pr\left\{ \int_{\Theta} [R(\theta) - R_n(\theta)]\, d\rho(\theta) \leq B_{n,C}(\pi, \rho, \varepsilon) \text{ for all } \rho \in \mathscr{P}_{\pi}(\Theta) \text{ simultaneously} \right\} \geq 1 - \varepsilon,$$

where

$$\begin{aligned} B_{n,C_n}(\pi, \rho, \varepsilon) = & \left[\frac{C_n}{1 - \exp(-C_n)} - 1 \right] \int_{\Theta} R_n(\theta)\, d\rho(\theta) \\ & + \frac{2U_{\max}}{n} \frac{1}{1 - \exp(-C_n)} \left[\log \frac{1}{\varepsilon} + D_{\mathrm{KL}}(\rho, \pi) \right]. \end{aligned}$$

When $R_n(\theta) > 0$, setting C_n proportional to $(n \int_{\Theta} R_n(\theta)\, d\rho(\theta))^{-1/2}$ yields the following best rate of the PAC bound $B_{n,C_n}(\pi, \rho, \varepsilon)$:

$$B_{n,C_n}(\pi, \rho, \varepsilon) = O_p\left(\sqrt{\frac{\int_{\Theta} R_n(\theta)\, d\rho(\theta)}{n}} \right).$$

It should be noted, however, that we cannot choose λ in C_n according to the data for the bounds in Theorem 2.3.2. We consider valid bounds when λ is data-dependent, for example when it is chosen via cross-validation in Theorems 2.3.4 and 2.3.5.

When $P(X,Y)$ and ρ are such that $R_n(a_{G,\rho}) = \int_{\Theta} R_n(\theta)\, d\rho(\theta)$ is very small, the PAC bound from Theorem 2.3.2(b) can be smaller than that in Theorem 2.3.2(a). For a given λ, Theorem 2.3.2(c) says that we can take the better of the two, without applying any union bound arguments that require a reduction in ε. On the other hand, Theorem 2.3.2(b) and (c) only provide upper bounds for $\int_{\Theta} R(\theta)\, d\rho(\theta)$ while 2.3.2(a) provides both an upper bound and a lower bound.

Note that Theorem 2.3.2 holds for all ρ simultaneously. Setting $\rho(\cdot)$ equal to $\hat{\rho}_{\lambda}(\cdot)$ in Theorem 2.3.2(a) and (c), we can obtain the following theorem.

Theorem 2.3.3 *Let Assumptions 2.2.1 – 2.2.4 hold. Then for $\varepsilon \in (0, 1]$ each of the following holds with probability at least $1 - \varepsilon$:*

(a)

$$\int_\Theta R(\theta)\,d\hat{\rho}_\lambda \le \int_\Theta R_n(\theta)\,d\hat{\rho}_\lambda + \frac{1}{\lambda}D_{\mathrm{KL}}(\hat{\rho}_\lambda,\pi) + \frac{1}{\lambda}\left[\frac{\lambda^2\left(K_\psi^2+\mu_\psi^2\right)}{n} + \log\frac{1}{\varepsilon}\right],$$

(b)

$$\left|\int_\Theta R(\theta)\,d\hat{\rho}_\lambda - \int_\Theta R_n(\theta)\,d\hat{\rho}_\lambda\right| \le \frac{1}{\lambda}D_{\mathrm{KL}}(\hat{\rho}_\lambda,\pi) + \frac{1}{\lambda}\left[\frac{\lambda^2\left(K_\psi^2+\mu_\psi^2\right)}{n} + \log\frac{2}{\varepsilon}\right]$$

(c)

$$\int_\Theta R(\theta)\,d\hat{\rho}_\lambda \le \min_{\rho\in\mathscr{P}_\pi(\Theta)}\left[\int_\Theta R(\theta)d\rho(\theta) + \frac{2}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right] + \frac{2}{\lambda}\left[\frac{\lambda^2\left(K_\psi^2+\mu_\psi^2\right)}{n} + \log\frac{2}{\varepsilon}\right].$$

If (2.27) *holds,* $(K_\psi^2+\mu_\psi^2)$ *can be replaced by* $U_{\max}^2/2$ *in (a)-(c).*

(d) When (2.27) *holds,*

$$\int_\Theta R(\theta)\,d\hat{\rho}_\lambda(\theta) \le \min\left\{U_{\lambda,\pi,\hat{\rho}_\lambda}(\varepsilon), U^{\mathscr{F}}_{\lambda,\pi,\hat{\rho}_\lambda}(\varepsilon)\right\},$$

where $U_{\lambda,\pi,\hat{\rho}_\lambda}(\varepsilon)$ *and* $U^{\mathscr{F}}_{\lambda,\pi,\hat{\rho}_\lambda}(\varepsilon)$ *are given by* (2.29) *and* (2.30) *with* ρ *set to* $\hat{\rho}_\lambda$*.*

Theorem 2.3.3(a) provides a PAC-Bayesian bound for the generalization error of the Gibbs method. When $U_{\max} < \infty$, choosing the rate-optimal $\lambda = \kappa\sqrt{n}$ for some constant $\kappa > 0$ gives us

$$\Pr\left\{\int_\Theta R(\theta)\,d\hat{\rho}_\lambda(\theta) \le \int_\Theta R_n(\theta)\,d\hat{\rho}_\lambda(\theta) + \frac{1}{\kappa\sqrt{n}}\left[D_{\mathrm{KL}}(\hat{\rho}_\lambda,\pi) + \log\frac{1}{\varepsilon}\right] + \frac{\kappa U_{\max}^2}{2\sqrt{n}}\right\} \ge 1-\varepsilon. \tag{2.31}$$

Therefore, the PAC generalization error decays to zero at the rate of $1/\sqrt{n}$.

Theorem 2.3.3(b) allows us to construct a $(1-\varepsilon)$ confidence interval $CI_{\lambda,\pi}(\varepsilon)$ for

$\int_{\Theta} R(\theta)\, d\hat{\rho}_{\lambda}(\theta)$:

$$CI_{\lambda,\pi}(\varepsilon) = \left[L_{\lambda,\pi}(\varepsilon), U_{\lambda,\pi}(\varepsilon)\right],$$

where

$$L_{\lambda,\pi}(\varepsilon) = \int_{\Theta} R_n(\theta)\, d\hat{\rho}_{\lambda} - \frac{1}{\lambda} D_{\mathrm{KL}}(\hat{\rho}_{\lambda}, \pi) - \frac{1}{\lambda}\left(\frac{\lambda^2\left(K_{\psi}^2 + \mu_{\psi}^2\right)}{n} + \log\frac{2}{\varepsilon}\right),$$
$$U_{\lambda,\pi}(\varepsilon) = \int_{\Theta} R_n(\theta)\, d\hat{\rho}_{\lambda} + \frac{1}{\lambda} D_{\mathrm{KL}}(\hat{\rho}_{\lambda}, \pi) + \frac{1}{\lambda}\left(\frac{\lambda^2\left(K_{\psi}^2 + \mu_{\psi}^2\right)}{n} + \log\frac{2}{\varepsilon}\right).$$

Let

$$U_{\lambda,\pi}^{\mathscr{F}}(\varepsilon) = \mathscr{F}_{n,\lambda}^{-1}\left(\int_{\Theta} R_n(\theta)\, d\hat{\rho}_{\lambda} + \frac{1}{\lambda} D_{\mathrm{KL}}(\hat{\rho}_{\lambda}, \pi) + \frac{1}{\lambda}\log\frac{2}{\varepsilon}\right).$$

Then the upper limit of $CI_{\lambda,\pi}(\varepsilon)$ can be replaced by $\min(U_{\lambda,\pi}(\varepsilon), U_{\lambda,\pi}^{\mathscr{F}}(\varepsilon))$, leading to a shorter confidence interval. This follows from a union bound argument as in the proof of Theorem 2.3.3(a). Note that $U_{\lambda,\pi}^{\mathscr{F}}(\varepsilon)$ above is equal to $U_{\lambda,\pi,\hat{\rho}_{\lambda}}^{\mathscr{F}}(\varepsilon/2)$ in equation (2.30). If there is a natural bound for $\int_{\Theta} R(\theta)\, d\hat{\rho}_{\lambda}(\theta)$, such as 0 for the lower bound or $2U_{\max}$ for the upper bound, we should make an obvious modification to the above interval.

Theorem 2.3.3(c) shows that the estimated probability measure $\hat{\rho}_{\lambda}(\cdot)$ strikes almost the best trade-off between the average risk $\int_{\Theta} R(\theta)\, d\rho(\theta)$ and the regularization term $\frac{2}{\lambda} D_{\mathrm{KL}}(\rho, \pi)$. The best trade-off that solves the minimization problem is given by the distribution $\rho_{\lambda R/2}$ with the following RN derivative

$$\frac{d\rho_{\lambda R/2}}{d\pi}(\theta) = \frac{\exp\left[-\frac{\lambda}{2} R(\theta)\right]}{\int_{\Theta} \exp\left[-\frac{\lambda}{2} R(\theta)\right] d\pi(\theta)}.$$

This follows from Corollary 2.2.1(a). Note that $R(\theta)$ is not feasible and is only known to an oracle. Hence, $\rho_{\lambda R/2}$ is not feasible and the bound in the theorem is an oracle-type risk bound.

Theorem 2.3.3(c) can be interpreted as selecting the best probability measure in $\mathscr{P}_{\pi}(\Theta)$.

Ideally, we select $\rho(\theta) \in \mathscr{P}_{\pi}(\Theta)$ to minimize the average risk $\int_{\Theta} R(\theta)\, d\rho(\theta)$. An oracle who knows $R(\theta)$ can solve for the best $\rho^{*}(\theta)$, namely, $\rho^{*} = \arg\min_{\rho \in \mathscr{P}_{\pi}(\Theta)} \int_{\Theta} R(\theta)\, d\rho(\theta)$. Not knowing $R(\theta)$, we replace it by the empirical estimator $R_n(\theta)$ and add a regularization term to the objective function. That is, we solve the optimization problem in (2.23). The selected $\hat{\rho}_{\lambda}$ can not be expected to be as good as ρ^{*}. However, Theorem 2.3.3(c) shows that it is almost as good as a second best oracle solution $\rho_{\lambda R/2}$.

In practice, λ will be chosen by cross-validation. However, cross validating λ inhibits the use of Theorems 2.3.2 and 2.3.3 for deriving risk bounds or confidence intervals. We mention two methods for dealing with this. First, we can employ an idea from Catoni (2007) for deriving bounds that do not rely on λ. This entails combining a union-bound argument with Theorem 2.3.2 and leads to the following theorem.

Theorem 2.3.4 *Let Assumptions 2.2.1 – 2.2.4 hold and let* $\alpha > 1$ *and* $\varepsilon \in (0,1]$. *Assume* (2.27) *holds. Each event below holds with probability at least* $1-\varepsilon$.

(a) For $s \in \{-1,1\}$ *and for all* $\rho \in \mathscr{P}_{\pi}(\Theta)$ *simultaneously,*

$$\int_{\Theta} s\,[R(\theta) - R_n(\theta)]\, d\rho(\theta) \leq \inf_{\lambda > 1} \left\{ \frac{\alpha}{\lambda} \left[\frac{\lambda^2 U_{\max}^2}{2n} + \log\frac{1}{\varepsilon} + D_{\mathrm{KL}}(\rho,\pi) + 2\log\frac{\log(\alpha^2\lambda)}{\log\alpha} \right] \right\}$$

(b) For $s \in \{-1,1\}$ *and any* $\tilde{\lambda} > 1$ *which may be chosen based on the sample,*

$$\int_{\Theta} s\,[R(\theta) - R_n(\theta)]\, d\hat{\rho}_{\tilde{\lambda}}(\theta) \leq \frac{\alpha}{\tilde{\lambda}} \left[\frac{\tilde{\lambda}^2 U_{\max}^2}{2n} + \log\frac{1}{\varepsilon} + D_{\mathrm{KL}}(\hat{\rho}_{\tilde{\lambda}},\pi) + 2\log\frac{\log\left(\alpha^2\tilde{\lambda}\right)}{\log\alpha} \right]$$

(c) Let

$$\mathscr{F}_{n,\lambda,\alpha}^{-1}(r) = 2U_{\max}\frac{1-\exp\left(-\frac{\lambda}{n}\cdot r\right)}{1-\exp\left(-\frac{\lambda}{\alpha n}\cdot 2U_{\max}\right)},$$

and define

$$\overline{U}_{\lambda,\pi,\rho,\alpha}(\varepsilon)=\int_{\Theta}R_n(\theta)d\rho(\theta)+\frac{\alpha}{\lambda}\left[\frac{\lambda^2U_{\max}^2}{2n}+\log\frac{1}{\varepsilon}+D_{\mathrm{KL}}(\rho,\pi)+2\log\frac{\log\left(\alpha^2\lambda\right)}{\log\alpha}\right],$$

$$\overline{U}^{\mathscr{F}}_{\lambda,\pi,\rho,\alpha}(\varepsilon)=\mathscr{F}^{-1}_{n,\lambda,\alpha}\left(\int_{\Theta}R_n\left(\theta\right)d\rho(\theta)+\frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)+\frac{1}{\lambda}\left[\log\frac{1}{\varepsilon}+2\log\frac{\log\left(\alpha^2\lambda\right)}{\log\alpha}\right]\right).$$

For all $\rho\in\mathscr{P}_{\pi}\left(\Theta\right)$ simultaneously,

$$\int_{\Theta}R(\theta)d\rho(\theta)\leq\inf_{\lambda>1}\left\{\min\left[\overline{U}_{\lambda,\pi,\rho,\alpha}(\varepsilon),\overline{U}^{\mathscr{F}}_{\lambda,\pi,\rho,\alpha}(\varepsilon)\right]\right\}.$$

(d) For any $\tilde{\lambda}>1$ that may be chosen based on the sample,

$$\int_{\Theta}R(\theta)d\hat{\rho}_{\tilde{\lambda}}(\theta)\leq\min\left[\overline{U}_{\tilde{\lambda},\pi,\hat{\rho}_{\tilde{\lambda}},\alpha}(\varepsilon),\overline{U}^{\mathscr{F}}_{\tilde{\lambda},\pi,\hat{\rho}_{\tilde{\lambda}},\alpha}(\varepsilon)\right].$$

Theorem 2.3.4 is stated for the case where $U_{\max}<\infty$, i.e., a setting where the bounds can be computed without knowledge of the DGP. However, the bounds in parts (a) and (b) have valid counterparts in the more general case where we would replace $U_{\max}^2/2$ by $K_{\psi}^2+\mu_{\psi}^2$. Following similar arguments to those producing the confidence interval $CI_{\lambda,\pi}\left(\varepsilon\right)$ after Theorem 2.3.3, a confidence interval for $\int_{\Theta}R(\theta)d\hat{\rho}_{\tilde{\lambda}}$ can be derived from Theorem (2.3.4) that is valid when $\hat{\rho}_{\tilde{\lambda}}$ is such that $\tilde{\lambda}$ is data-dependent. Note that in parts (a) and (c) the infimum is taken over all $\lambda>1$. The condition that $\lambda>1$ is fairly reasonable in relation to the bounds that motivate the decision rules. To see this, in the bounded utility setting, suppose that $U:\{-1,1\}^2\times\mathscr{X}\to[-U_{\max},U_{\max}]$ is replaced with the normalized utility $\tilde{U}=U/(2U_{\max})$, which of course does not alter the underlying preferences. Then $\tilde{U}_{\max}=\sup_{a,y,x}|\tilde{U}(a,y,x)|=1/2$, so that the point-forecast loss based on this utility function satisfies $0\leq\tilde{\ell}(\theta,y,x)\leq1$. With this normalization, any observed loss is then a percentage of the largest possible loss rather than relying on potentially arbitrary utils. With this normalization, for any $0 < \lambda \leq 1$, both the bounds in parts (a) and (b) of Theorem 2.3.2 are such that the right-hand side is trivial (i.e., it is at least 1) whenever $\varepsilon < \exp(-1)$. Focusing on $\lambda > 1$ restricts attention to values for which confidence in the bounds is more reasonable.

A second method for obtaining bounds or confidence intervals when $\tilde{\lambda}$ is data-dependent is to build from bounds in the literature where the PAC-Bayesian analysis does not utilize this temperature parameter. For example, the following result is also obtained in Maurer (2004) and Germain et al. (2015). While these authors do not explicitly consider loss functions that vary with X, some results there carry through when the utility function is bounded.

Lemma 2.3.2 *Let Assumptions 2.2.1 – 2.2.4 hold and assume that* (2.27) *holds. Let*

$$D(r_1, r_2) = \frac{n}{\lambda}\left[\mathrm{kl}\left(\frac{r_2}{2U_{\max}}, \frac{r_1}{2U_{\max}}\right)\right], \text{ where } \mathrm{kl}(a,b) = a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}.$$

Then, for $\lambda > 0$, condition (2.25) *in Theorem 2.3.1 holds with*

$$f(\lambda, n) = \log \xi(n), \text{ where } \xi(n) := \sum_{k=1}^{n}\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}.$$

That $\mathrm{kl}(\cdot,\cdot)$ is convex follows from Theorem 2.7.2 of Cover and Thomas (2006) and we adopt the convention that $0\log 0 = 0$, $a\log\frac{a}{0} = \infty$ if $a > 0$ and $0\log\frac{0}{0} = 0$. Note that $\mathrm{kl}(a,b)$ is the KL-divergence between two Bernoulli random variables with success probabilities a and b respectively. It can be shown (c.f. Lemma 19 in Germain et al. (2015) and references therein) that $\sqrt{n} \leq \xi(n) \leq 2\sqrt{n}$. Theorem 2.3.1 combined with Lemma 2.3.2 produce the first part of the following theorem. The second part follows from an application of Pinsker's inequality, $2(a-b)^2 \leq \mathrm{kl}(a,b)$.

Theorem 2.3.5 *Let Assumptions 2.2.1 – 2.2.4 hold and assume that* (2.27) *holds. For $\varepsilon > 0$, each of the following holds with probability at least $1-\varepsilon$.*

(a) for all $\rho \in \mathscr{P}_\pi(\Theta)$ simultaneously,

$$\mathrm{kl}\left(\frac{R_n(a_{G,\rho})}{2U_{\max}}, \frac{R\left(a_{G,\rho}\right)}{2U_{\max}}\right) \leq \frac{1}{n}\left[\log \xi(n) + \log\frac{1}{\varepsilon} + D_{\mathrm{KL}}(\rho,\pi)\right].$$

(b) for all $\rho \in \mathscr{P}_\pi(\Theta)$ simultaneously,

$$\left|\int_\Theta R(\theta)d\rho(\theta) - \int_\Theta R_n(\theta)d\rho(\theta)\right| \leq 2U_{\max}\sqrt{\frac{1}{2n}\left(\log\xi(n) + \log\frac{1}{\varepsilon} + D_{\mathrm{KL}}(\rho,\pi)\right)}.$$

As discussed in Germain et al. (2015), (a) is a slight improvement over similar bounds that have arisen in earlier PAC-Bayesian literature. One option to derive a bound for

$$\int_\Theta R_n(\theta)d\hat{\rho}_{\tilde{\lambda}}(\theta)$$

is to solve the inequality in (a) numerically. As the bounds in Theorem 2.3.5 do not depend on λ and are valid for any $\rho \in \mathscr{P}_\pi(\Theta)$, they produces bounds for $\hat{\rho}_{\tilde{\lambda}}$ when $\tilde{\lambda}$ is data dependent.

Lastly, the generalization bounds for the loss function can be used to obtain generalization bounds for the utility function directly. To this end, denote

$$\mathbb{U}_n(\theta) = \frac{1}{n}\sum_{i=1}^n [U(a(X,\theta),Y,X)] = \frac{1}{n}\sum_{i=1}^n U(Y_i,Y_i,X_i) - R_n(\theta)$$
$$\text{and } \mathbb{U}(\theta) = E[U(a(X,\theta),Y,X)] = EU(Y,Y,X) - R(\theta).$$

Also let

$$B_U = U_{\max}\sqrt{\frac{2\log\frac{2}{\varepsilon}}{n}}.$$

Then we have the following corollary of earlier bounds for $\int_\Theta [R(\theta) - R_n(\theta)]d\hat{\rho}(\theta)$.

Corollary 2.3.1 *For $\varepsilon > 0$, let $\hat{\rho}$ be a probability distribution over Θ and let $B_R(\hat{\rho})$ be a high* *probability (at least* $1-\varepsilon/2$*) bound for* $\int_\Theta [R(\theta)-R_n(\theta)]d\hat{\rho}(\theta)$*, i.e.* $B_R(\hat{\rho})$ *satisfies*

$$\Pr\left\{\int_\Theta [R(\theta)-R_n(\theta)]\, d\hat{\rho}(\theta) \leq B_R(\hat{\rho})\right\} \geq 1-\frac{\varepsilon}{2}.$$

Then

$$\Pr\left(\int_\Theta \mathbb{U}(\theta)d\hat{\rho}(\theta) \geq \int_\Theta \mathbb{U}_n(\theta)d\hat{\rho}(\theta) - (B_U + B_R(\hat{\rho}))\right) \geq 1-\varepsilon.$$

For example, if we are considering decision rules using $\hat{\rho}_{\tilde{\lambda}}$ with data dependent $\tilde{\lambda} > 1$, under the assumptions of Theorem 2.3.4(a) and for $\alpha > 1$ we can take

$$B_R(\hat{\rho}_{\tilde{\lambda}}) = \frac{\alpha}{\tilde{\lambda}}\left[\frac{\tilde{\lambda}^2 U_{\max}^2}{2n} + \log\frac{1}{\varepsilon} + D_{\mathrm{KL}}(\hat{\rho}_{\tilde{\lambda}}, \pi) + 2\log\frac{\log\left(\alpha^2\tilde{\lambda}\right)}{\log\alpha}\right].$$

2.3.2 Linear Decision Rules in the Utility Setting

By definition, the estimator $\hat{\rho}_\lambda$ solves

$$\hat{\rho}_\lambda := \underset{\rho\in\mathscr{P}_\pi(\Theta)}{\arg\min}\left[\int_\Theta R_n(\theta)\, d\rho(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right].$$

The distribution is not standard and must be approximated by numerical methods such as MCMC or tempered SMC (the latter is discussed in Section 2.4). Here, we consider a restrictive class of posteriors from a parametric family. In particular, we consider the case that both ρ and π are normal. Specifically, we assume that under π

$$\theta = \left(\theta_1, \theta_2, ..., \theta_q\right)' \sim N(\mu_\pi, \Sigma_\pi),$$

and under ρ

$$\theta = \left(\theta_1, \theta_2, ..., \theta_q\right)' \sim N(\mu_\rho, \Sigma_\rho),$$

where μ_π and μ_ρ are the mean vectors and Σ_π and Σ_ρ are the covariance matrices.

Lemma 2.3.3 *The KL divergence between* $\rho : N(\mu_\rho, \Sigma_\rho)$ *and* $\pi : N(\mu_\pi, \Sigma_\pi)$ *on* $\mathbb{R}^q$ *is*

$$D_{\text{KL}}(\rho, \pi) = \frac{1}{2}(\mu_\rho - \mu_\pi)' \Sigma_\pi^{-1} (\mu_\rho - \mu_\pi) + \frac{1}{2}\left[\text{tr}\left(\Sigma_\rho \Sigma_\pi^{-1}\right) - q\right] - \frac{1}{2}\log\frac{\det\left(\Sigma_\rho\right)}{\det\left(\Sigma_\pi\right)}.$$

We further assume that $\mathscr{R}_\Theta$ is described by (2.14) and that for $x \in \mathscr{X}$,

$$m(x, \theta) = \sum_{j=1}^{q} \phi_j(x)\theta_j = \phi(x)'\theta,\ \theta \in \mathbb{R}^q \tag{2.32}$$

for some set of feature transformations $\{\phi_1(x), \ldots., \phi_q(x)\}$ where $\phi_j(x) : \mathscr{X} \to \mathbb{R}$. For example, $\left\{\phi_1(x), \ldots, \phi_q(x)\right\}$ can consist of transforms of the observable variables using any set of basis functions. Another case of interest would be the setting where $\mathscr{R}_\Theta$ is specified by

$$\mathscr{R}_\Theta = \left\{a(x, \theta) = \text{sign}\left(\phi(x)'\theta\right) : \theta \in \mathbb{R}^q\right\} \tag{2.33}$$

This is analogous to the setting of Germain et al. (2009) in the 0/1 loss setting. For example, one could take $\{\phi_1(x), \ldots., \phi_q(x)\}$ to be a set of decision stumps, with a fixed number of stumps and predetermined thresholds for each component of $x \in \mathbb{R}^d$. We focus on (2.32) below, but the results are easily adjusted to the setting of (2.33), simply drop the term $c(x)$.

Before proceeding, we note that the majority vote or Bayes method in this setting takes a particularly convenient form. For any fixed X, note that under $\theta \sim N(\mu_\rho, \Sigma_\rho)$ we have

$$\phi(X)'\theta - c(X) \sim N\left(\phi(X)'\mu_\rho - c(X), \phi(X)'\Sigma_\rho\phi(X)\right),$$

and therefore it follows that

$$E_{\theta\sim\rho}\text{sign}\left[\phi(X)'\theta - c(X)\right] = 2\Phi\left(\frac{\phi(X)'\mu_\rho - c(X)}{\sqrt{\phi(X)'\Sigma_\rho\phi(X)}}\right) - 1.$$

Hence the majority vote takes the form

$$
\begin{aligned}
a_{B,\rho}(X) &= \operatorname{sign}\left\{E_{\theta\sim\rho}\operatorname{sign}\left[\phi(X)'\theta - c(X)\right]\right\} \\
&= \operatorname{sign}\left\{2\Phi\left(\frac{\phi(X)'\mu_\rho - c(X)}{\sqrt{\phi(X)'\Sigma_\rho\phi(X)}}\right) - 1\right\} = \operatorname{sign}\left[\phi(X)'\mu_\rho - c(X)\right].
\end{aligned}
$$

That is, the decision rule in this case is straightforward to calculate and depends directly on a linear combination of a set of mappings from $\mathscr{X}$ to $\mathbb{R}$. Additionally, we will utilize the following lemma.

Lemma 2.3.4 *Under the normal prior and posterior setting described above,*

$$
\int_\Theta R_n(\theta)\, d\rho(\theta) = \frac{1}{n}\sum_{i=1}^n \psi(X_i,Y_i)\Phi\left(-\frac{V\left(X_i,Y_i,\mu_\rho\right)}{\sqrt{\phi(X_i)'\Sigma_\rho\phi(X_i)}}\right).
$$

where

$$
V\left(X_i,Y_i,\mu_\rho\right) = Y_i\left[\phi(X_i)'\mu_\rho - c(X_i)\right].
$$

Given Lemma 2.3.4, the minimization problem then reduces to the following problem:

$$
\left(\hat{\mu}_\rho, \hat{\Sigma}_\rho\right) := \arg\min_{\mu_\rho,\Sigma_\rho}\left\{\frac{1}{n}\sum_{i=1}^n \psi(X_i,Y_i)\,\Phi\left(-\frac{V\left(X_i,Y_i,\mu_\rho\right)}{\sqrt{\phi(X_i)'\Sigma_\rho\phi(X_i)}}\right) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right\}.
$$

When $\mu_\pi = 0$, $\Sigma_\pi = diag\left(\sigma^2_{\pi,j}\right)$, and $\Sigma_\rho = diag\left(\sigma^2_{\rho,j}\right)$, we have

$$
\begin{aligned}
D_{\mathrm{KL}}(\rho,\pi) &= \frac{1}{2}\sum_{j=1}^q \frac{\mu^2_{\rho,j}}{\sigma^2_{\pi,j}} + \frac{1}{2}\left[\sum_{j=1}^q \frac{\sigma^2_{\rho,j}}{\sigma^2_{\pi,j}} - q\right] - \frac{1}{2}\sum_{j=1}^q \log\frac{\sigma^2_{\rho,j}}{\sigma^2_{\pi,j}} \\
&= \frac{1}{2}\left[\sum_{j=1}^q \frac{\mu^2_{\rho,j}}{\sigma^2_{\pi,j}} + \sum_{j=1}^q\left(\frac{\sigma^2_{\rho,j}}{\sigma^2_{\pi,j}} - \log\frac{\sigma^2_{\rho,j}}{\sigma^2_{\pi,j}}\right) - q\right],
\end{aligned}
$$

and the minimization problem becomes

$$\left(\hat{\mu}_\rho, \hat{\sigma}^2_\rho\right) := \arg\min_{\mu_\rho,\sigma^2_\rho} \frac{1}{n}\sum_{i=1}^{n} \psi(X_i,Y_i)\,\Phi\left(-\frac{V\left(X_i,Y_i,\mu_\rho\right)}{\sqrt{\sum_{j=1}^{q}\sigma^2_{\rho,j}\phi_j(X_i)^2}}\right) + \frac{1}{2\lambda}\sum_{j=1}^{q}\left(\frac{\mu^2_{\rho,j}}{\sigma^2_{\pi,j}} + \frac{\sigma^2_{\rho,j}}{\sigma^2_{\pi,j}} - \log\frac{\sigma^2_{\rho,j}}{\sigma^2_{\pi,j}}\right). \tag{2.34}$$

Given the estimator $\hat{\sigma}^2_\rho = (\hat{\sigma}^2_{\rho,1},...,\hat{\sigma}^2_{\rho,q})$, we have

$$\hat{\mu}_\rho := \arg\min_{\mu_\rho}\left\{\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)\,\Phi\left(-\frac{V\left(X_i,Y_i,\mu_\rho\right)}{\sqrt{\sum_{j=1}^{q}\hat{\sigma}^2_{\rho,j}\phi_j(X_i)^2}}\right) + \frac{1}{2\lambda}\sum_{j=1}^{q}\frac{\mu^2_{\rho,j}}{\sigma^2_{\pi,j}}\right\}.$$

The first term can be regarded as the empirical loss function for μ_ρ, and the second term is a weighted L_2 regularizer. If all of $\left\{\hat{\sigma}^2_{\rho,j}\right\}$ converge to zero, which is expected, then

$$\Phi\left(-\frac{V\left(X_i,Y_i,\mu_\rho\right)}{\sqrt{\sum_{j=1}^{q}\hat{\sigma}^2_{\rho,j}\phi_j(X_i)^2}}\right) \approx 1\left\{V\left(X_i,Y_i,\mu_\rho\right) < 0\right\}.$$

In addition to the weighted L_2 regularization, the PAC-Bayesian approach, therefore, also replaces the indicator

$$1\left\{V\left(X,Y,\mu_\rho\right) < 0\right\},$$

which is not smooth, by a smooth function

$$\Phi(-V\left(X,Y,\mu_\rho\right)/h),$$

for a small h. Smoothing and regularization are two built-in features of the PAC-Bayesian approach.

In the econometric literature, smoothing has been proposed to overcome the technical difficulties behind the maximum score estimator. See, for example, Horowitz (1992). In instrumental variable quantile regressions where an indicator function is present in the criterion function, Kaplan and Sun (2017) discuss several benefits of smoothing, including variance reduction and computational convenience. The PAC-Bayesian approach provides another justification for smoothing.

Lastly, we consider a particular form of the restrictive model considered here that will be utilized in the simulation section and is easier to implement. When $\Sigma_\pi = \Sigma_\rho = I_M$ and $\mu_\pi = 0$, we seek only to estimate μ_ρ and the optimization problem is now equivalent to

$$\hat{\mu}_\rho = \arg\min_{\mu_\rho} \frac{\lambda}{n} \sum_{i=1}^{n} \psi(X_i,Y_i)\Phi\left(-\frac{V\left(X_i,Y_i,\mu_\rho\right)}{||\phi(X_i)||}\right) + \frac{1}{2}||\mu_\rho||^2. \tag{2.35}$$

The resulting decision rule is given by

$$a(x,\hat{\mu}_\rho) = \text{sign}[\phi(x)'\hat{\mu}_\rho - c(x)].$$

Here λ is a hyperparameter we will choose via cross-validation. Alternatively, the version corresponding to the model class in (2.33) would drop the term $c(X_i)$, and is just a weighted version of the model derived in Germain et al. (2009), where the only difference in the objective function above is the weighting term $\psi(X_i,Y_i)$. Germain et al. (2009) utilizes the 0/1 based loss version of this model with $\{\phi_1(X),\ldots,\phi_M(X)\}$ taken as a set of weak learning decision stumps and show that the estimator performs competitively against AdaBoost in terms of misclassification rates on several real world data sets.

Note that (2.35) exhibits similarities with the soft-margin support vector machine, which selects $\hat{\mu}_{svm}$ to minimize the objective function

$$C\sum_{i=1}^{n}\left[1 - Y_i\phi(X_i)'\mu_\rho\right]_+ + \frac{1}{2}||\mu_\rho||^2$$

for some constant $C > 0$ and has classification rule $a(x,\hat{\mu}_{svm}) = \text{sign}[\phi(x)'\hat{\mu}_{svm}]$. In the restrictive PAC-Bayesian objective function in (2.35), a bounded and smooth “sigmoid” loss replaces the hinge loss of the SVM and now the terms in the objective function are weighted by $\psi(X_i,Y_i)$, the missed payoff from an incorrect decision.

2.3.3 PAC-Bayesian Multi-Model Aggregation

Here we consider the situation where there are multiple binary decision model classes of interest. Section 2.3.1 is general enough to encompass this setting with only some notational changes and reinterpretations. Here we detail the changes in the model space, prior specification, and posterior distribution, and present some implications relevant to implementation in this setting.

Suppose there are now K models indexed by $k = 1, 2, ..., K$. Let $\theta_{(k)} \in \mathbb{R}^{q_k}$ be the parameter vector for model k. The number of parameters q_k can be different for a different model. For example, different decision boundaries may consist of a different subset of covariates, and the size of the subset can be different. Denote $\theta = \left(k, \theta_{(k)}\right)$. The first component of θ signifies the model class, and the second component signifies the model parameter given the model class in the first component. The parameter space for θ is

$$\Theta = \cup_{k=1}^{K} \left(k \times \Theta_{(k)}\right),$$

where $\Theta_{(k)}$ is the parameter space for $\theta_{(k)}$. Given $\theta = \left(k, \theta_{(k)}\right) \in \Theta$, the action function, now denoted by $a_{(k)}(x, \theta_{(k)})$, maps the covariate space $\mathscr{X}$ to a binary action. The single model setting in Section 2.3.1 can be regarded as a special case here with $k = K = 1$.

As before, we equip Θ with the standard σ-algebra denoted by $\mathscr{B}_\theta$. PAC-Bayesian learning for model aggregation works in the same way as before. We need to specify a "prior" distribution π over the (model, parameter)-pairs $\left\{\left(k, \theta_{(k)}\right)\right\}$ in the measurable space $(\Theta, \mathscr{B}_\theta)$ and then use the performances of different pairs to update π to obtain an "evidence-based" distribution. The final decision rule involves aggregating the actions of all (model, parameter)-pairs using the evidence-based distribution.

To specify a distribution π over Θ, we first specify the distribution $\pi(k)$ over the model classes $k = 1, ..., K$ and then specify the distribution $\pi\left(\theta_{(k)}|k\right)$ over $\theta_{(k)} \in \Theta_{(k)}$ given the model class k. Let $\mathscr{K}^{\circ}$ be a subset of $\mathscr{K} := \{1, 2, ..., K\}$ and $\Theta^{\circ}_{(k)}$ be a measurable subset of $\Theta_{(k)}$. Then $\Theta^{\circ} = \cup_{k \in \mathscr{K}^{\circ}}(k \times \Theta^{\circ}_{(k)})$ is a measurable subset of Θ. Based on $\pi(k)$ and $\pi\left(\theta_{(k)}|k\right)$, $\pi(\Theta^{\circ})$ is defined as

$$\pi(\Theta^{\circ}) = \sum_{k \in \mathscr{K}^{\circ}} \left[\pi(k) \cdot \int_{\Theta^{\circ}_{(k)}} d\pi\left(\theta_{(k)}|k\right)\right].$$

With some abuse of notation[2], we write the measure π as

$$\pi(\theta) := \pi\left(\left(k, \theta_{(k)}\right)\right) = \pi(k)\,\pi\left(\theta_{(k)}|k\right) \text{ for } \theta = \left(k, \theta_{(k)}\right). \tag{2.36}$$

This gives a general characterization of any distribution on $(\Theta, \mathscr{B}_{\theta})$.

Given a $\pi \in \mathscr{P}(\Theta)$, we denote the family of all distributions on $(\Theta, \mathscr{B}_{\theta})$ that is absolutely continuous with respect to π as $\mathscr{P}_{\pi}(\Theta)$. The evidence-based distribution we consider will belong to $\mathscr{P}_{\pi}(\Theta)$. For any $\rho \in \mathscr{P}_{\pi}(\Theta)$, define the Kullback–Leibler divergence between ρ and π as

$$D_{\mathrm{KL}}(\rho, \pi) = \sum_{k=1}^{K} \left\{\int_{\Theta_{(k)}} \log\left[\frac{\rho(k)}{\pi(k)} \cdot \frac{d\rho\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)}\right] d\rho\left(\theta_{(k)}|k\right)\right\} \rho(k).$$

This is the same definition as before but is tailored to the model aggregation setting with new interpretations of $\theta \in \Theta$ and the distribution over Θ.

Let $\mathscr{M}(\Theta)$ be the set of measurable functions on $(\Theta, \mathscr{B}_{\theta})$ and

$$\mathscr{M}^{\pi}_{b}(\Theta) = \left\{A : A(\cdot,\cdot) \in \mathscr{M}(\Theta) \text{ and } \sum_{k=1}^{K}\left[\int_{\Theta_{(k)}} \exp\left(A\left(k, \theta_{(k)}\right)\right) d\pi\left(\theta_{(k)}|k\right)\right] \pi(k) < \infty\right\},$$

which is a subset of $\mathscr{M}(\Theta)$ that has a finite exponential moment under π. In this setting, Lemma 2.2.2 can be stated as follows.

Lemma 2.3.5 *For $\pi \in \mathscr{P}(\Theta)$ and $A \in \mathscr{M}(\Theta)$ such that $-A \in \mathscr{M}^{\pi}_{b}(\Theta)$, let $\rho_{A,\pi} \in \mathscr{P}_{\pi}(\Theta)$ be*

[2] Here the meaning of $\pi(\cdot)$ depends on the argument supplied. We could write $\pi(\theta)$ as $\pi_{\theta}(\theta)$, $\pi(k)$ as $\pi_{k}(k)$ and $\pi\left(\theta_{(k)}|k\right)$ as $\pi_{\theta_{(k)}|k}\left(\theta_{(k)}|k\right)$ but we opt for a more economical notation. This should not cause any confusion.

the probability measure on Θ *defined by*

$$\rho_{A,\pi}(\theta) = \rho_{A,\pi}(k) \cdot \rho_{A,\pi}\left(\theta_{(k)}|k\right), \textit{ for } \theta = \left(k, \theta_{(k)}\right),$$

where

$$\rho_{A,\pi}(k) = \frac{\pi(k)\,\nu_A(k)}{\sum_{j=1}^{K}\pi(j)\,\nu_A(j)},$$

$$\frac{d\rho_{A,\pi}\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)} = \frac{\exp\left(-A\left(k,\theta_{(k)}\right)\right)}{\nu_A(k)},$$

and

$$\nu_A(k) = \int_{\Theta_{(k)}} \exp\left(-A\left(k,\tilde{\theta}_{(k)}\right)\right) d\pi\left(\tilde{\theta}_{(k)}|k\right).$$

That is, for any measurable set $\Theta^\circ = \cup_{k\in\mathscr{K}^\circ}(k \times \Theta^\circ_{(k)}) \subseteq \Theta$,

$$\rho_{A,\pi}(\Theta^\circ) = \sum_{k\in\mathscr{K}^\circ}\left[\rho_{A,\pi}(k) \cdot \int_{\Theta^\circ_{(k)}} d\rho_{A,\pi}\left(\theta_{(k)}|k\right)\right].$$

Then, for any probability measure $\rho \in \mathscr{P}_\pi(\Theta)$ *we have*

$$\begin{aligned}&\log\left[\sum_{k=1}^{K}\pi(k)\,\nu_A(k)\right]\\ &= -\left\{D_{\mathrm{KL}}(\rho,\pi) + \sum_{k=1}^{K}\left[\int_{\Theta_{(k)}} A\left(k,\theta_{(k)}\right) d\rho\left(\theta_{(k)}|k\right)\right]\rho(k)\right\} + D_{\mathrm{KL}}\left(\rho,\rho_{A,\pi}\right).\end{aligned}$$

Note that $\log\left[\sum_{k=1}^{K}\pi(k)\,\nu_A(k)\right]$ does not depend on ρ. It follows from Lemma 2.3.5 that

$$\begin{aligned}&\arg\min_{\rho\in\mathscr{P}_\pi(\Theta)}\left\{D_{\mathrm{KL}}(\rho,\pi) + \sum_{k=1}^{K}\left[\int_{\Theta_{(k)}} A\left(k,\theta_{(k)}\right) d\rho\left(\theta_{(k)}|k\right)\right]\rho(k)\right\}\\ &= \arg\min_{\rho\in\mathscr{P}_\pi(\Theta)} D_{\mathrm{KL}}\left(\rho,\rho_{A,\pi}\right) = \rho_{A,\pi}. \end{aligned} \tag{2.37}$$

With the above details for the model aggregation setting, we can return to the optimization problem similar to that in (2.20). Let $R_n\left(k,\theta_{(k)}\right)$ be the empirical risk under model k with parameter $\theta_{(k)}$:

$$R_n\left(k,\theta_{(k)}\right)=\frac{1}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)1\left\{Y_i\neq a_{(k)}(X_i,\theta_{(k)})\right\}.$$

We now solve

$$\min_{\rho\in\mathscr{P}_\pi(\Theta)}\left[E_{\left(k,\theta_{(k)}\right)\sim\rho}\left[R_n(k,\theta_{(k)})\right]+\frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right]. \tag{2.38}$$

To characterize the solution to the above minimization problem, we define the data-dependent measure on $\mathscr{K}$ as

$$\hat{\rho}_\lambda\left(k\right)=\frac{\pi\left(k\right)\hat{v}_\lambda\left(k\right)}{\sum_{j=1}^{K}\pi\left(j\right)\hat{v}_\lambda\left(j\right)} \tag{2.39}$$

and the data-dependent measure $\hat{\rho}_\lambda\left(\theta_{(k)}|k\right)$ on $\Theta_{(k)}$ in terms of its RN derivative with respect to $\pi\left(\theta_{(k)}|k\right)$ as

$$\frac{d\hat{\rho}_\lambda\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)}=\frac{\exp\left(-\lambda R_n\left(k,\theta_{(k)}\right)\right)}{\hat{v}_\lambda\left(k\right)},\ k=1,\ldots,K,$$

where

$$\hat{v}_\lambda\left(k\right)=\int_{\Theta_{(k)}}\exp\left(-\lambda R_n\left(k,\theta_{(k)}\right)\right)d\pi\left(\theta_{(k)}|k\right).$$

Based on $\hat{\rho}_\lambda\left(k\right)$ and $\hat{\rho}_\lambda\left(\theta_{(k)}|k\right)$, we form the data-dependent measure $\hat{\rho}_\lambda\left(\theta\right)\in\mathscr{P}\left(\Theta\right)$ according to

$$\hat{\rho}_\lambda\left(\Theta^\circ\right)=\sum_{\ell\in\mathscr{L}^\circ}\left[\hat{\rho}_\lambda\left(k\right)\cdot\int_{\Theta^\circ_{(k)}}d\hat{\rho}_\lambda\left(\theta_{(k)}|k\right)\right]. \tag{2.40}$$

This is our evidence-based distribution over (model, parameter)-pairs.

Letting

$$A\left(k,\theta_{(k)}\right)=\lambda R_n\left(k,\theta_{(k)}\right),$$

Lemma 2.3.5 and equation (2.37) thereafter show that $\hat{\rho}_\lambda\left(\theta\right)$ solves the problem in (2.38).

One approach to evaluate decision rules based on $\hat{\rho}_\lambda\left(\theta\right)$ is to simulate this distribution via reversible jump MCMC. Alternatively, as we consider in this paper, when the majority vote classifier is the object of interest, the form of $\hat{\rho}_\lambda\left(\theta\right)$ is amenable to the SMC method. For the single model class setting, the SMC approach is described in Section 2.4. One benefit of the SMC approach is that the procedure based on a single model class is easily adapted to the multiple model class setting. When the form of $\pi\left(\theta_{(k)}|k\right)$ does not depend on the choice for $\pi(k)$, this can reduce the computational burden if one is interested in choosing the prior component $\pi(k)$ over $\mathscr{K}$ from a set of potential distributions over $\mathscr{K}$ via cross-validation.

To see this, note that the majority vote (or, Bayesian) decision rule based on $\hat{\rho}_\lambda$ is

$$a_{B,\hat{\rho}_\lambda}(x) = \operatorname{sign}\left\{E_{(k,\theta_{(k)})\sim\hat{\rho}_\lambda} a_{(k)}(x,\theta_{(k)})\right\} = \operatorname{sign}\left\{\sum_{k=1}^{K} \hat{\rho}_\lambda\left(k\right)\hat{a}_{(k)}\left(x\right)\right\}, \qquad (2.41)$$

where

$$\hat{a}_{(k)}\left(x\right) := \int_{\Theta_{(k)}} a_{(k)}(x,\theta_{(k)}) d\rho_{\hat{\lambda}}\left(\theta_{(k)}|k\right).$$

For a single model class $\mathscr{R}_{\Theta_{(k)}}$ with a given $\pi\left(\theta_{(k)}|k\right)$, under general conditions the SMC procedure produces accurate estimators for $\hat{a}_{(k)}\left(x\right)$ and $\hat{v}_\lambda(k)$, both of which depend only on $\pi\left(\theta_{(k)}|k\right)$. These objects can be computed separately for each $k \in \mathscr{K}$ according to the single model class SMC procedure. Then, for a given $\pi(k)$ over $\mathscr{K}$, (2.39) can be used to construct $\hat{\rho}_\lambda(k)$ and the majority vote rule is computed via (2.41). If one is interested in cross-validating the choice of $\pi(k)$ from some set of distributions on $\mathscr{K}$ and the distributions $\pi\left(\theta_{(k)}|k\right)$ do not depend on $\pi(k)$ for $k \in \mathscr{K}$, then the objects $\hat{a}_{(k)}\left(x\right)$ and $\hat{v}_\lambda(k)$ need only be computed once per cross-validation sample. This is in contrast to running a reversible jump MCMC procedure for each choice of $\pi(k)$ and can be beneficial when the number of decision model classes is not very large. If the number of model classes was very large, say, in an explanatory variable selection setting where the total number of explanatory variables is greater than the sample size, then an alternative computational strategy would be needed (to avoid running the SMC procedure independently for each model class). See, for example, Guedj (2013) for a discussion of PAC-Bayesian analysis and implementation for binary outcomes in such a setting.

2.4 Implementation

Here we consider implementation choices and describe some settings of the computational procedures that are applied in our simulations in Section 2.5. In Section 2.4.1 we discuss prior choices and consider examples for $\mathscr{R}_\Theta$. The $\mathscr{R}_\Theta$ considered center on decision models similar to those in Su (2020) and Elliott and Lieli (2013), some of which are used in our simulations. However, it should not be too difficult to make adjustments if a different model class is desired. In Section 2.4.2 we discuss the calculation of $\hat{\mu}_\rho$ in (2.35) associated with the linear decision rule discussed at the end of Section 2.3.2. We also outline an implementation of the SMC algorithm of Del Moral et al. (2006) in our setting in Section 2.4.2.

2.4.1 Model and Prior Choices

First we consider two specifications for $\mathscr{R}_\Theta$ of the form in (2.14) that are also considered in Su (2020). These consist of specifying a functional form for $m(x,\theta) \in \mathscr{M}_\Theta$ and the associated parameter space Θ. Then we consider potential choices for the prior probability distribution π. The $\mathscr{R}_\Theta$ specifications allow for $m(x,\theta)$ to be fairly general and are appropriate for a setting where the number of explanatory variables d is not large relative to the sample size n. If d is larger than n, the choices of function class and prior utilized in Guedj (2013) (Chapter 3) would be an option; an MCMC-based approach would be more appropriate in such a setting rather than the SMC procedure in Section 2.4.2.

In many empirical applications, we have a nondecreasing collection of parameterized function classes $\{\mathscr{M}_{\Theta_{(k)}}\}_{k=1}^K$ for $K \in \mathbb{N}$ where $\mathscr{M}_{\Theta_{(i)}} \subset \mathscr{M}_{\Theta_{(j)}}$ for $i < j$. In a single model class setting, we can take $\mathscr{M}_\Theta$ to be $\mathscr{M}_{\Theta_{(k)}}$ for some $k \in \mathscr{K} = \{1,\ldots,K\}$ with parameter space $\Theta = \Theta_{(k)}$. In the multiple model class setting of Section 2.3.3, we can take $\mathscr{M}_\Theta = \cup_{k=1}^K \mathscr{M}_{\Theta_{(k)}}$ with parameter space $\Theta = \cup_{k=1}^K (k \times \Theta_{(k)})$. While the inclusion of the model class k as a component of the parameter θ may seem redundant when the model classes are nested, it simplifies the prior specification and allows for generalization when the model classes are not nested.

Example 1 *We consider polynomial transformations on $\mathscr{X}$ of order at most $k \in \mathscr{K}$. For $\mathscr{X} \subset \mathbb{R}^d$, the polynomial transformation of order at most k will have $q_k = \binom{d+k}{k}$ parameters, and it is defined as*

$$\mathscr{M}^{\text{poly}}_{\Theta_{(k)}} = \left\{ m(x,\theta) = \sum_{j=1}^{q_k} \theta_j \phi_j(x), \;\; \theta_{(k)} = (\theta_1, \dots, \theta_{q_k}) \in \mathbb{R}^{q_k} \right\},$$

where the summation is over all monomials $\phi_j(x) = \prod_{\ell=1}^{d} x_\ell^{p_{j\ell}}$ with $\sum_{\ell=1}^{d} p_{j\ell} \leq k$ and $p_{j\ell} \in \mathbb{N} \cup \{0\}$. The parameter space associated with $\mathscr{M}^{\text{poly}}_{\Theta_{(k)}}$ is $\Theta_{(k)} = \mathbb{R}^{q_k}$.

Example 2 *Define $\Lambda(v) = (1 + \exp(-v))^{-1}$. With the same parameter set $\Theta_{(k)} = \mathbb{R}^{q_k}$ as in Example 1, define the function space*

$$\mathscr{M}^{\text{logistic}}_{\Theta_{(k)}} = \left\{ m(x,\theta) = \Lambda\left(f(x,\theta)\right) : f(x,\theta) \in \mathscr{M}^{\text{poly}}_{\Theta_{(k)}} \right\}.$$

Now we consider some options for specifying the prior. First consider when $\mathscr{M}_\Theta$ and Θ correspond to a single model class, i.e., $\mathscr{M}_\Theta = \mathscr{M}_{\Theta_{(k)}}$ for some fixed $k \in \mathscr{K}$. In cases where it is reasonable to bound the parameter space Θ (for example, one could possibly replace $\Theta = \mathbb{R}^{q_k}$ with a bounded subset of $\mathbb{R}^{q_k}$ given some knowledge about the distribution of $P(X,Y)$), a uniform prior over Θ is a potential choice. When $\Theta = \mathbb{R}^{q_k}$, another choice is a multivariate normal prior over Θ, for example, $N(0, \sigma_\pi^2 I_{q_k})$ for some $\sigma_\pi^2 > 0$. In the multiple model class setting with varying class complexity, a general strategy is to choose π that puts increasingly less weight on regions of the parameter space that are increasingly more complex. A prior that puts relatively more weight on very complex regions of the parameter space will tend to result in larger $D_{\text{KL}}(\rho, \pi)$ terms in the bounds of Section 2.3.1 particularly as λ increases.

In our simulations, we use the following formulation for π in the $\Theta = \cup_{k=1}^{K}(k \times \Theta_{(k)})$ setting. We specify π as in (2.36), taking $\pi(\theta_k|k)$ to be the $N(0, \sigma_\pi^2 I_{q_k})$ distribution for $k \in \mathscr{K}$ with a fixed $\sigma_\pi^2 > 0$. To specify the model-class component $\pi(k)$ of the prior, in addition to simpler schemes such as equal weighting, one choice we consider is to set

$$\pi(k) = \frac{\exp\left(-\eta\xi\left(k,n\right)\right)}{z_\eta}, \ \ z_\eta = \sum_{k=1}^{K} \exp(-\eta\xi(k,n)), \tag{2.42}$$

where $\eta \geq 0$ and $\xi(k,n) : \mathscr{K} \times \mathbb{N} \to \mathbb{R}_+$ is some measure of the complexity of model class k. Potential building blocks for $\xi(k,n)$ in the form of distribution-free model complexity measurements are as follows. For $k \in \mathscr{K}$, define $\mathscr{M}_{k,c} \equiv \{x \mapsto \operatorname{sign}(m(x,\theta) - c(x)) : m \in \mathscr{M}_{\Theta_{(k)}}\}$ and denote the growth function[3] of $\mathscr{M}_{k,c}$ by $\Pi_{k,c}(\cdot)$. Let $\psi_c(k,n)$ denote an upper bound for $\Pi_{k,c}(n)$ and $V_{k,c}$ denote an upper bound for the VC-dimension[4] of $\mathscr{M}_{k,c}$. That is, $\psi_c(k,n)$ upper bounds the maximum number of distinct ways that $\left\{\operatorname{sign}(m(x,\theta_{(k)}) - c(x)), \theta_{(k)} \in \Theta_k\right\}$ can classify any set of points in $\mathscr{X}^n$ while $V_{k,c}$ upper bounds the size of the largest sample that $\mathscr{M}_{k,c}$ could classify without error. To penalizes complexity, $\xi(k,n)$ can be taken to be an increasing function of $V_{k,c}$, $\psi_c(k,n)$, or both. In our simulations, we consider taking

$$\xi(k,n) = \sqrt{\log V_{k,c}}, \tag{2.43}$$

and also cross-validate η in (2.42) from a finite set of values.

Remark 1 below contains additional details regarding $V_{k,c}$ and $\psi_c(k,n)$ for Examples 1 and 2. These points are also noted in Su (2020); we refer the reader to their Section 3.1 and the references therein for additional discussion.

Remark 1 *When $\mathscr{M}_{\Theta_{(k)}}$ is specified as a vector space of real valued functions, the VC-dimension of $\mathscr{M}_{k,c}$ is given by the dimension of $\mathscr{M}_{\Theta_{(k)}}$ (c.f. Theorem 3.5 of Anthony and Bartlett (2009)). In particular, $\mathscr{M}^{\text{poly}}_{\Theta_{(k)}}$ in Example 1 has dimension $\binom{d+k}{k}$ when $\mathscr{X}$ does not contain dummy variables, and so we can take $V_{k,c} = \binom{d+k}{k}$. For Example 2 with $\mathscr{M}^{\text{logistic}}_{\Theta_{(k)}}$, Su (2020) shows that the VC-dimension of $\mathscr{M}_{k,c} = \{x \mapsto \operatorname{sign}(m(x,\theta) - c(x)) : m \in \mathscr{M}^{\text{logistic}}_{\Theta_{(k)}}\}$ can be bounded above*

[3] For a collection $\mathscr{H}$ of functions from $\mathscr{X}$ to $\{-1,1\}$, $\Pi_{\mathscr{H}} : \mathbb{N} \to \mathbb{N}$ is defined by $\Pi_{\mathscr{H}}(\ell) = \max_{(x_1,\ldots,x_\ell)\in\mathscr{X}^\ell} |\{(h(x_1),\ldots,h(x_\ell)) : h \in \mathscr{H}\}|$

[4] The VC-dimension of $\mathscr{M}_{k,c}$ is the largest integer ℓ such that $\Pi_{k,c}(\ell) = 2^\ell$.

by $\binom{d+k}{k}+1$, and hence we can take $V_{k,c}=\binom{d+k}{k}+1$. Regarding $\psi_c(k,n)$, if $V_{k,c}$ bounds the VC-dimension of $\mathscr{M}_{k,c}$, it follows from Theorems 3.5 and 3.6 in Anthony and Bartlett (2009) that $\Pi_{k,c}(n)$ can be upper bounded by

$$\psi_c(k,n)=\begin{cases}2^n, & \text{if } n\leq V_{k,c}\\ \left(\frac{en}{V_{k,c}}\right)^{V_{k,c}}, & \text{if } n>V_{k,c}.\end{cases}$$

2.4.2 Implementation for Methods in Section 2.3

In Section 2.5, our simulations evaluate the majority vote or Bayes method with the Gibbs posterior $\hat{\rho}_\lambda$ in Definition 2.2.1 and also with the linear decision rule in Section 2.3.2 associated with the optimization problem in (2.35). Here we first detail our approach to computing $\hat{\mu}_\rho$ in the latter case and then outline the SMC approach applied to implement $\hat{\rho}_\lambda$ in the former and more general case. We address only the single model class setting here. The discussion in Section 2.3.3 highlights how this can be adapted to the multiple model class setting for $\hat{\rho}_\lambda$.

First, we make a computational adjustment so that the choice of the hyperparameter λ is invariant to the units of measurement of the utility function. For $\mathscr{P}^*\subseteq\mathscr{P}_\pi(\Theta)$ and $M\in\mathbb{N}$, note that cross-validating $\lambda\in\{\lambda_1,\ldots,\lambda_M\}$ among distributions in

$$\underset{\rho\in\mathscr{P}^*}{\arg\min}\left[\int_\Theta R_n(\theta)d\rho(\theta)+\frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right]$$

is equivalent to cross-validating $\lambda\in\{\lambda_1\bar{\psi},\ldots,\lambda_M\bar{\psi}\}$ among distributions in

$$\underset{\rho\in\mathscr{P}^*}{\arg\min}\left[\int_\Theta \bar{R}_n(\theta)d\rho(\theta)+\frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right], \tag{2.44}$$

where $\bar{R}_n(\theta)=R_n(\theta)/\bar{\psi}$, and $\bar{\psi}=n^{-1}\sum_{i=1}^n\psi(X_i,Y_i)$. We work with the adjusted minimization problem in (2.44). In the general setting where the Gibbs posterior $\hat{\rho}_\lambda$ is considered, $\mathscr{P}^*$ is $\mathscr{P}_\pi(\Theta)$ whereas in the linear decision setting associated with the optimization problem in (2.35),

$\mathscr{P}^*$ is the set of normal distributions over Θ with an identity covariance matrix (and π is the standard normal distribution).

To compute $\hat{\mu}_\rho$ in the setting of Section 2.3.2 discussed above, we follow a similar strategy to that of Germain et al. (2009) who analyze the 0/1-loss version of this problem. Incorporating the adjustment in (2.44) into the objective in (2.35), the optimization problem is now

$$\hat{\mu}_\rho = \arg\min_{\mu_\rho} \frac{\lambda}{n} \sum_{i=1}^{n} \frac{\psi(X_i, Y_i)}{\bar{\psi}} \Phi\left(- \frac{V\left(X_i, Y_i, \mu_\rho\right)}{||\phi(X_i)||} \right) + \frac{1}{2}||\mu_\rho||^2.$$

The gradient of the objective function above with respect to μ_ρ is given by

$$-\frac{\lambda}{n} \sum_{i=1}^{n} \frac{\psi(X_i, Y_i)}{\bar{\psi}} \dot{\Phi}\left(\frac{Y_i\left[\phi(X_i)\mu_\rho - c(X_i)\right]}{||\phi(X_i)||} \right) \frac{Y_i \phi(X_i)}{||\phi(X_i)||} + \mu_\rho,$$

where $\dot{\Phi}$ denotes the standard normal probability density function. For a given value of λ, $\hat{\mu}_\rho$ is calculated by gradient descent. As there can be multiple local minima, we tried 15 random starting points when $\lambda/n \leq 10$ and 100 random starting points when $\lambda/n > 10$. We performed 5-fold cross-validation to select $\lambda \in \{2^0, 2^1, \ldots, 2^{18}\}$. The discussion and references in Alquier et al. (2016) suggest alternative implementation methods. These can be useful for the more general settings in Section 2.3.2, for example when the covariance matrix Σ_ρ is not set to the identity matrix.

To implement the majority vote rule based on $\hat{\rho}_\lambda$ in Definition 2.2.1, now with $R_n(\theta)$ replaced by $\bar{R}_n(\theta)$, we utilize the tempering SMC procedure of Del Moral et al. (2006). While MCMC is a typical choice for simulating from $\hat{\rho}_\lambda$, recently Ridgway et al. (2014) and Alquier et al. (2016) have highlighted the usefulness of the SMC procedure in various PAC-Bayesian settings. One benefit is that each run of the procedure produces a sample from each member of a set of Gibbs posterior distributions corresponding increasing λ values. This can ease the computational burden of cross-validation.

To touch on a few elements of the tempering SMC algorithm in our setting, assume

$\Theta = \mathbb{R}^q$ for some $q \in \mathbb{N}$ and that we are able to sample from a prior probability distribution π over Θ. It is assumed that there is an increasing temperature ladder

$$0 = \lambda_0 < \lambda_1 < \cdots < \lambda_T, \ T \in \mathbb{N}.$$

$\{\lambda_t\}_{t=0}^T$ here is not generally the same set that was considered for cross-validation in the earlier procedure for the linear decision rule. The temperature ladder is intended to be such that as λ_t increases, the corresponding distributions $\hat{\rho}_{\lambda_t}$ progress gradually from $\pi = \hat{\rho}_{\lambda_0}$ to distributions $\hat{\rho}_{\lambda_t}$ with higher values of λ_t that are of greater interest. For each $t = 0, \ldots, T$, the SMC algorithm produces a set of weighted samples, $\{W_t^{(i)}, \theta_t^{(i)}\}_{i=1}^N$ with $W_t^{(i)} > 0$ and $\sum_{i=1}^N W_t^{(i)} = 1$, of size N and a scaling factor estimate $\hat{Z}_t$. The set of parameter draws $\{\theta_t^{(i)}\}_{i=1}^N$ are referred to as particles (there are N weighted particles for each t). SMC combines MCMC moves with sequential importance sampling. This produces weighted particles that emulate, in terms of computing expectations, samples from the probability distributions $\hat{\rho}_{\lambda_t}$ associated with the densities

$$\frac{d\hat{\rho}_{\lambda_t}}{d\pi}(\theta) = \frac{\exp\left[-\lambda_t \bar{R}_n(\theta)\right]}{Z_t}, \ Z_t = \int_\Theta \exp\left[-\lambda_t \bar{R}_n(\theta)\right] d\pi(\theta), \ t = 0, 1, \ldots, T.$$

Under general conditions, for a $\hat{\rho}_{\lambda_T}$-integrable function $\varphi : \Theta \to \mathbb{R}$,

$$\sum_{i=1}^N W_T^{(i)} \varphi\left(\theta_T^{(i)}\right) \overset{a.s.}{\to} E_{\theta \sim \hat{\rho}_{\lambda_T}} \varphi(\theta),$$

as $N \to \infty$ while $\hat{Z}_T$ is consistent for Z_T. In our setting we are interested in $\varphi(\theta) = a(x, \theta)$ where $a(x, \theta) \in \mathscr{R}_\Theta$, enabling us to compute the key ingredient to the majority vote decision rule. For additional details regarding the SMC procedure and its applications, we refer to Del Moral et al. (2006) and Jasra et al. (2007).

The SMC algorithm we apply in Section 2.5 is detailed below. We set the input parameters τ_{ESS} and N there equal to $1/2$ and 1000, respectively. For the $\{\lambda_t\}_{t=1}^T$ input, we adopt the piece-wise linear structure utilized in the simulations of Del Moral et al. (2006) and Jasra et al. (2007) with $T = 320$ and $\lambda_T = 1600$. In particular, the first 20% of steps increase uniformly from 0 to 0.15×1600 (i.e. $\lambda_j = (j/64) \times 240$ for $j = 1, \ldots, 64$), the next 40% of steps increase uniformly from 240 to 0.4×1600 (i.e. $\lambda_j = 240 + (j/128) \times 400$ for $j = 65, \ldots, 192$), and the last 40% of steps increase uniformly from 640 to 1600 (i.e. $\lambda_j = 640 + (j/128) \times 960$ for $j = 193, \ldots, 320$). In practice, it may be beneficial to consider higher (or lower) values of λ_T and or include a greater number of steps (higher T value). Depending on the data generating process, higher values of λ_T can push some components such as $\hat{Z}_t$ close to machine epsilon for t near T. One can experiment a little to check that the temperature range doesn't appear to be limited unnecessarily and if increasing the number of steps improves performance in cross-validation samples. Alternatives to the piece-wise linear ladder design are discussed in Del Moral et al. (2006) and Jasra et al. (2007). Additionally, the SMC algorithm requires a resampling step. We utilize systematic resampling, which is also outlined below. Additional algorithm choices and cross-validation points are detailed below the algorithm descriptions.

Tempering SMC Algorithm

Input N (number of particles), $\tau_{\text{ESS}} \in (0,1)$ (ESS threshold), $\{\lambda_t\}_{t=1}^T$ (temperature ladder with $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_T$).

Output $\{W_t^{(i)}, \theta_t^{(i)}\}_{i=1}^N$ for $t = 0, \ldots, T$, $\{\hat{Z}_t\}_{t=1}^T$.

Step 1: initialization

- Set $t \leftarrow 0$, $\hat{Z}_0 \leftarrow 1$. For $i = 1, \ldots, N$, draw $\theta_0^{(i)} \sim \pi$ and set $W_0^{(i)} \leftarrow 1/N$.

Iterate steps 2 and 3

Step 2: Resampling

- If

$$\left\{\sum_{i=1}^{N}\left(W_t^{(i)}\right)^2\right\}^{-1} < \tau_{\mathrm{ESS}}N,$$

resample $\left\{W_t^{(i)},\theta_t^{(i)}\right\}_{i=1}^N$ yielding equally weighted resampled particles $\left\{\frac{1}{N},\overline{\theta}_t^{(i)}\right\}_{i=1}^N$ and set $\left\{W_t^{(i)},\theta_t^{(i)}\right\}_{i=1}^N \leftarrow \left\{\frac{1}{N},\overline{\theta}_t^{(i)}\right\}_{i=1}^N$. Otherwise, leave $\left\{W_t^{(i)},\theta_t^{(i)}\right\}_{i=1}^N$ unaltered.

Step 3: Sampling

- Set $t \leftarrow t+1$; if $t = T+1$, stop.

- For $i = 1,\dots,N$, draw $\theta_t^{(i)} \sim K_t(\theta_{t-1}^{(i)},\cdot)$, where K_t is an MCMC kernel with invariant distribution ρ_{λ_t}, and evaluate the unnormalized importance weights

$$\omega_t^{(i)}\left(\theta_{t-1}^{(i)}\right) = \exp\left[-(\lambda_t-\lambda_{t-1})\bar{R}_n\left(\theta_{t-1}^{(i)}\right)\right].$$

- For $i = 1,\dots,N$, set

$$W_t^{(i)} \leftarrow \frac{W_{t-1}^{(i)}\omega_t\left(\theta_{t-1}^{(i)}\right)}{\sum_{j=1}^N W_{t-1}^{(j)}\omega_t\left(\theta_{t-1}^{(j)}\right)},\ \hat{Z}_t \leftarrow \hat{Z}_{t-1}\times\left\{\sum_{i=1}^N W_{t-1}^{(i)}\omega_t\left(\theta_{t-1}^{(i)}\right)\right\}.$$

Resampling Algorithm (systematic resampling):

Input A set of (normalized) weights and associated particles, $\left\{W_t^{(i)},\theta_t^{(i)}\right\}_{i=1}^N$ for some $t \in \{0,\dots,T\}$.

Output Resampled particles for equal weighting, $\left\{\overline{\theta}_t^{(i)}\right\}_{i=1}^N$

- Draw $u \sim U\left[0,\frac{1}{N}\right]$.

- Compute cumulative weights $C^{(i)} = \sum_{m=1}^i W_t^{(m)}$ for $i = 1,...,N$.

```
- Set $m \leftarrow 1$.
- For $i = 1 : N$
        While $u < C^{(i)}$ do $\overline{\theta}_t^{(m)} \leftarrow \theta_t^{(i)}$.
        $m \leftarrow m+1$, and $u \leftarrow u + 1/N$.
  End For
```

For the MCMC kernel in the sampling step of the SMC algorithm, we use a Gaussian random-walk Metropolis kernel with covariance matrix proportional to the empirical covariance matrix of the current set of particles. We scale the empirical covariance of the step t particles by $1/t$ which produced produced reasonable acceptance rates in the first simulated training set across the various simulation setups. The priors utilized for the majority vote associated with the Gibbs posterior in our simulations are described in Section 2.5 below. We use 5-fold cross-validation to select λ from λ_t values for which $t > 25$.

2.5 Simulation Study

To investigate the performance of the utility-based PAC-Bayesian decision rules, we consider two data generating processes and two sets of preferences, one set with each DGP. We utilize the same simulation design as Elliott and Lieli (2013) and Su (2020). The DGPs and the associated sets of preferences are as follows.

DGP 1: $\mathscr{X} = [-2.5, 2.5]$, $X \sim 5 \times \text{Beta}(1, 1.3) - 2.5$, and $P(x) = \Lambda(-0.5X + 0.2X^3)$ where $\Lambda(\cdot)$ is the logistic function described in Example 2 and recall $P(x)$ is defined in (2.2).

- Preference 1: $b(x) = 20$ and $c(x) = 0.5$.
- Preference 2: $b(x) = 20$ and $c(x) = 0.5 + 0.025X$.

DGP 2: $\mathscr{X} = [-3.5, 3.5]^2$, covariates X_1 and X_2 are each uniformly distributed on $[-3.5, 3.5]$ and are independent of one another, and $P(x_1, x_2) = \Lambda(Q(1.5x_1 + 1.5x_2))$ where $Q(v) = (1.5 - 0.1v)\exp\{-(0.25v + 0.1v^2 - 0.04v^3)\}$.

- Preference 3: $b((x_1, x_2)) = 20$ and $c((x_1, x_2)) = 0.75$.
- Preference 4: $b((x_1, x_2)) = 20 + 40 \cdot 1\{|x_1 + x_2| < 1.5\}$ and $c((x_1, x_2)) = 0.75$.

To evaluate the performance of a decision rule, we compute, by Monte Carlo simulation, the ratio of its expected utility to the expected utility of the optimal decision in (2.9) if $P(x)$ were known. This metric is intuitive as utility has no natural unit, however the ratio changes when a constant is added to the utility function. In Elliott and Lieli (2013) and Su (2020), this is dealt with by choosing some normalization of the utility function. We follow the same normalization and Monte Carlo setup as Su (2020), so that our simulation results can be compared directly to theirs. Noting that

$$U(a, y, x) = \frac{1}{4}b(x)\,[y + 1 - 2c(x)]\,a + \frac{1}{4}b(x)\,[y + 1 - 2c(x)] + U(-1, y, x),$$

Su (2020) normalizes the utility function by setting $U(-1, y, x) = -0.25b(x)[y + 1 - 2c(x)]$ for all $x \in \mathscr{X}$ and multiplying the utility function by 4. For any decision rule $a_n(x) : \mathscr{X} \to \{-1, 1\}$, this results in the following measurement that he calls the generalized expected utility,

$$S(a_n) = E\{b(X)[Y + 1 - 2c(X)]a_n(X)\}.$$

With this normalization, denote

$$a^*(x) = \text{sign}[P(x) - c(x)],\ x \in \mathscr{X},$$

i.e., a^* is the optimal forecast rule. Then define the relative generalized expected utility (RGEU) of any decision rule a_n by

$$\text{RGEU}(a_n) \equiv \frac{E\left[S(a_n(X))\right]}{S(a^*(X))}.$$

As noted in Su (2020), the RGEU of the decision rule a_n can be approximated by simulation as

$$\text{RGEU}(a_n) = E\left[\frac{S(a)}{S(a^*)}\right] \simeq \frac{1}{\mathscr{S}} \sum_{j=1}^{\mathscr{S}} \frac{S_{\ell,j}\left(a|\mathscr{D}_{n,j}\right)}{S_{\ell,j}(a^*)}.$$

Here, $S_{\ell,j}(a_n|\mathscr{D}_{n,j})$ is the jth out of sample empirical utility with training sample size ℓ of the decision rule a_n, which is estimated on the jth training sample $\mathscr{D}_{n,j}$ with training sample size n. $S_{\ell,j}(a^*)$ is the jth out-of-sample empirical utility with training sample size ℓ of a^*, and $\mathscr{S}$ is the number of simulation replications. Still following Su (2020), we take $n \in \{500, 1000\}$, $\ell = 5000$, and $\mathscr{S} = 500$.

We compare the following models. Firstly, we consider maximum likelihood estimators, which are denoted by ML in Tables 1 and 2. For $k = 1, 2, 3$, the maximum likelihood estimator presumes a logistic model linear in the polynomial transformations of the $\mathscr{X}$ up to order k. Secondly, we consider the maximum utility estimator of Elliott and Lieli (2013) (denoted MU); it is presumed that $m(x, \theta)$ belongs to the class of polynomial transformations of $\mathscr{X}$ for $k = 1, 2, 3$ for these decision rules. Hence the ML estimator is correctly specified for $P(X)$ when $k = 3$ for DGP 1. Thirdly, we consider one of the best performing (in this simulation design) model selection procedures from Su (2020), based on the simulated maximal discrepancy penalty. This is a penalized version of the MU models here (selecting the best k among $k = 1, 2, 3$). This model is denoted MU-SMD. Fourthly, we consider the linear PAC-Bayesian model associated with (2.35) from Section 2.3.2 when the posterior is also constrained to be normal with identity covariance matrix. Here we take $\{\phi_1, \ldots, \phi_{q_3}\}$ to consist of the polynomial transformations of $\mathscr{X}$ up to order 3. We normalize the data (using training sample mean and standard deviation) as is common with SVM. This model is denoted PB-NP (NP for normal posterior). Lastly, we consider the non-constrained PAC-Bayesian method whereby the decision rule is the majority vote associated with the Gibbs posterior $\hat{\rho}_{\lambda}$ in Definition 2.2.1. In this case, we consider the multiple model class setting of Section 2.3.3. For the model classes, we use consider 3 classes of polynomial transformations on $\mathscr{X}$ of orders $k \in \{1,2,3\} = \mathscr{K}$ as specified in Example 1. We cross-validate λ according to the temperature ladder described in Section 2.4.2 and take $\pi(\theta_{(k)}|k)$ to be $N(0, 4I_{q_k})$ for each k. These decision rules are denoted PB-GP (GP for Gibbs posterior). To specify $\pi(k)$ for $k \in \mathscr{K}$, we evaluate three choices. First, we take $\pi(k) = 1/3$ for $k = 1,2,3$; this is denoted EQ. Second, we take $\pi(k) = q_k/(\sum_{j=1}^{3} q_j)$ where q_k is defined in Example 1 and denotes the number of parameters associated with model class k; this is denoted NP. Third, we utilize the weights in (2.43) and cross-validate $\tau \in \{2^{-2}, 2^{-1}, \ldots, 2^{3}\}$; this prior choice is denoted CV in the tables.

The simulation results are presented in Tables 1 and 2 after the Conclusion. The utility-based PAC-Bayesian decision models PB-GP and PB-NP perform very well, achieving higher RGEU than the MU and MU-SMD decision rules across all preferences and DGPs. The margin of the improvements is often sizable. Only the ML rule with a correctly specified DGP (ML with $k = 3$ for DGP 1) outperforms the BP- models. However, whenever the ML procedure is misspecified, it mostly performs quite poorly relative to all the utility-based methods. This performance further deteriorates when the preferences vary with the covariates as they do for Preferences 2 and 4. As shown in Elliott and Lieli (2013), the cubic MU ($k = 3$) is correctly specified in both the DGP 1 and DGP 2 settings. However, it is also observed there that MU can be prone to overfitting and aided by model selection procedures. Nonetheless, the PB- models outperform against the MU-SMD procedure in this simulation setting as well.

The restricted PAC-Bayesian decision model, PB-NP, performs slightly worse than the general version associated with the Gibbs posterior, PB-GP in most settings. However, the margin between the PB-NP and PB-GP models is not always very sizable. This may suggest that the restricted model can stand on its own, particularly when the sample size is larger or when there is a set $\{\phi_j(x)\}_{j=1}^{q}$ of interest that could be difficult to work into a more general Gibbs posterior setting. For example, when $\{\phi_j(x)\}_{j=1}^{q}$ is a larger set of weak learners as in Germain et al. (2009), the setting of Section 2.3.2 may be easier to implement. Lastly, we did not observe much of an impact on the RGEU from cross-validating the choice of τ in the prior $\pi(k)$.

2.6 Conclusion

An asymmetric payoff structure is often a salient feature of economic decision making problems. For the binary decision/forecast problem where the decision maker faces asymmetric payoffs that vary with observable variables, we propose a PAC-Bayesian approach. We show that many key elements of the PAC-Bayesian classification literature can be extended to accommodate this setting, deriving high probability training sample bounds and oracle inequalities that suggest decision rules of interest. The decision rules perform very well against alternatives methods in Monte Carlo experiments, allow for flexible functional decision rule forms, allow for valid training-sample risk bounds and confidence interval computation, and can take advantage of Bayesian estimation machinery.

Chapter 2 contains material being prepared for submission for academic publication. It is joint work with Yixiao Sun. The dissertation author is a primary author of this material.

Table 1. Relative generalized expected utility, $n = 500$

DGP 1	$P(x) = \Lambda(-0.5x + 0.2x^3)$					
Preference	$b(x) = 20$, $c(x) = 0.5$			$b(x) = 20$, $c(x) = 0.5 + 0.025x$		
π class weighting:	EQ	NP	CV	EQ	NP	CV
PB-GP	81.74	81.28	80.72	81.84	81.82	80.41
PB-NP	74.21			77.13		
MU-SMD	65.54			58.87		
Poly. order:	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
ML	34.16	29.57	93.09	9.13	10.92	94.37
MU	51.02	52.85	65.74	32.40	43.80	53.26
DGP 2	$P(x) = \Lambda(Q(1.5x_1 + 1.5x_2))$, $Q(v) = \frac{(1.5-0.1v)}{\exp(0.25v+0.1v^2-0.04v^3)}$					
Preference	$b(x) = 20$, $c(x) = 0.75$			$b(x) = 20 + 1\|x_1 + x_2\| < 1.5$, $c(x) = 0.75$		
π class weighting:	EQ	NP	CV	EQ	NP	CV
PB-GP	72.81	72.62	72.49	61.77	61.65	61.40
PB-NP	69.75			56.45		
MU-SMD	68.81			52.84		
Poly. order:	$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
ML	60.17	58.75	59.48	29.52	27.87	33.81
MU	66.71	51.87	67.69	48.35	33.14	51.47

Note: The MATLAB packages *glmfit* and *simulannealbnd* with default settings for each algorithm were used to compute the ML and MU models. The code implementing the ML, MU and MU-SMD models was provided by the author of Su (2020).

Table 2. Relative generalized expected utility, $n = 1000$

DGP 1	$P(x)=\Lambda(-0.5x+0.2x^3)$					
Preference	$b(x)=20$, $c(x)=0.5$			$b(x)=20$, $c(x)=0.5+0.025x$		
π class weighting:	EQ	NP	CV	EQ	NP	CV
PB-GP	87.73	87.86	87.47	90.81	90.65	90.43
PB-NP	81.52			88.83		
MU-SMD	70.75			67.30		
Poly. order:	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
ML	30.92	30.03	96.97	7.12	6.26	97.42
MU	53.24	58.19	69.50	36.91	49.13	60.41
DGP 2	$P(x)=\Lambda(Q(1.5x_1+1.5x_2))$, $Q(v)=\frac{(1.5-0.1v)}{\exp(0.25v+0.1v^2-0.04v^3)}$					
Preference	$b(x)=20$, $c(x)=0.75$			$b(x)=20+1\lvert x_1+x_2\rvert<1.5$, $c(x)=0.75$		
π class weighting:	EQ	NP	CV	EQ	NP	CV
PB-GP	78.61	78.46	78.25	70.09	70.18	69.86
PB-NP	73.72			63.75		
MU-SMD	71.94			59.72		
Poly. order:	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
ML	58.71	57.36	59.40	27.16	24.15	31.14
MU	69.97	58.14	70.81	54.92	39.24	55.97

Note: The MATLAB packages *glmfit* and *simulannealbnd* with default settings for each algorithm were used to compute the ML and MU models. The code implementing the ML, MU and MU-SMD models was provided by the author of Su (2020).

Appendices

2.A Appendix of Proofs for Chapter 2

2.A.1 Proofs for Section 2.2

Proof of Lemma 2.2.1. First we will show that for any $(x,y) \in \mathscr{X} \times \{-1,1\}$,

$$\psi(x,y)\,1\left\{y \neq a_{B,\rho}(x)\right\} \leq 2E_{\theta\sim\rho}\psi(x,y)1\left\{y \neq a(x,\theta)\right\}. \tag{2.45}$$

To show this, note that $\psi(x,y) > 0$ by Assumption 2.2.1 (i) and the fact that $\psi(x,y) = U(1,1,x) - U(-1,1,x)$ when $y=1$ and $\psi(x,y) = U(-1,-1,x) - U(1,-1,x)$ when $y=-1$. Therefore, (2.45) holds when $y = a_{B,\rho}(x) = \text{sign}\left\{E_{\theta\sim\rho}a(x,\theta)\right\}$ as then the left hand side is zero. When $y \neq a_{B,\rho}(x)$, this implies that $y \cdot E_{\theta\sim\rho}a(x,\theta) \leq 0$. Therefore in the alternative case when $y \neq a_{B,\rho}(x)$,

$$\begin{aligned}
\psi(x,y)1\left\{y \neq a_{B,\rho}\right\} &= \psi(x,y) \\
&\leq \psi(x,y)\left\{1 - yE_{\theta\sim\rho}a(x,\theta)\right\} \\
&= 2E_{\theta\sim\rho}\psi(x,y)\frac{1}{2}\left\{1 - y\cdot a(x,\theta)\right\} \\
&= 2E_{\theta\sim\rho}\psi(x,y)1\left\{y \neq a(x,\theta)\right\}.
\end{aligned}$$

This shows that (2.45) holds. By (2.45) and the monotonicity of expectation,

$$E_{X,Y\sim P(X,Y)}\psi(X,Y)\,1\left\{Y \neq a_{B,\rho}(X)\right\} \leq 2E_{X,Y\sim P(X,Y)}E_{\theta\sim\rho}\psi(X,Y)1\left\{Y \neq a(X,\theta)\right\}.$$

The statement of Lemma 2.2.1 then follows from an application of Fubini's theorem. ■

Proof of Lemma 2.2.2. By definition, we have

$$
\begin{aligned}
& D_{\mathrm{KL}}\left(\rho, \rho_{A,\pi}\right) \\
& = \int_{\Theta} \log\left[\frac{d\rho}{d\rho_{A,\pi}}(\theta)\right] d\rho(\theta) \\
& = \int_{\Theta} \log\left\{\frac{d\rho}{d\pi}(\theta)\left[\frac{d\rho_{A,\pi}}{d\pi}(\theta)\right]^{-1}\right\} d\rho(\theta) \\
& = \int_{\Theta} \left[\log\frac{d\rho}{d\pi}(\theta) - \log\frac{\exp\left(-A(\theta)\right)}{\int_{\Theta}\exp\left(-A\left(\tilde{\theta}\right)\right) d\pi\left(\tilde{\theta}\right)}\right] d\rho(\theta) \\
& = \int_{\Theta} A(\theta)\, d\rho(\theta) + \int_{\Theta} \log\left[\int_{\Theta}\exp\left(-A\left(\tilde{\theta}\right)\right) d\pi\left(\tilde{\theta}\right)\right] d\rho(\theta) + \int_{\Theta}\left[\log\frac{d\rho}{d\pi}(\theta)\right] d\rho(\theta) \\
& = \int_{\Theta} A(\theta)\, d\rho(\theta) + \log\left[\int_{\Theta}\exp(-A(\theta))\, d\pi(\theta)\right] + \int_{\Theta}\left[\log\frac{d\rho}{d\pi}(\theta)\right] d\rho(\theta) \\
& = \int_{\Theta} A(\theta)\, d\rho(\theta) + \log\left[\int_{\Theta}\exp(-A(\theta))\, d\pi(\theta)\right] + D_{\mathrm{KL}}(\rho, \pi).
\end{aligned}
$$

Hence,

$$
\log\left[\int_{\Theta}\exp(-A(\theta))\, d\pi(\theta)\right] = -\left[\int_{\Theta} A(\theta)\, d\rho(\theta) + D_{\mathrm{KL}}(\rho,\pi)\right] + D_{\mathrm{KL}}\left(\rho, \rho_{A,\pi}\right).
$$

■

Proof of Corollary 2.2.1.

Part (a). Since $\rho_{A,\pi} = \arg\min_{\rho \in \mathscr{P}_\pi(\Theta)} D_{\mathrm{KL}}\left(\rho, \rho_{A,\pi}\right)$ and the left hand side of (2.21) does not depend on ρ, we have

$$
\begin{aligned}
\rho_{A,\pi} &= \arg\max_{\rho \in \mathscr{P}_\pi(\Theta)} -\left[\int_{\Theta} A(\theta)\, d\rho(\theta) + D_{\mathrm{KL}}(\rho,\pi)\right] \\
&= \arg\min_{\rho \in \mathscr{P}_\pi(\Theta)} \left[\int_{\Theta} A(\theta)\, d\rho(\theta) + D_{\mathrm{KL}}(\rho,\pi)\right].
\end{aligned}
$$

By (1.29), we then have

$$
\begin{aligned}
&\min_{\rho\in\mathscr{P}_{\pi}(\Theta)}\left[\int_{\Theta}A(\theta)\,d\rho(\theta)+D_{\mathrm{KL}}(\rho,\pi)\right]\\
&=\int_{\Theta}A(\theta)\,d\rho_{A,\pi}(\theta)+D_{\mathrm{KL}}\left(\rho_{A,\pi},\pi\right)\\
&=-\log\left[\int_{\Theta}\exp(-A(\theta))\,d\pi(\theta)\right].
\end{aligned}
$$

Part (b). Taking $A=-\mathscr{A}$ in Lemma 2.2.2, we obtain that for any probability measure $\rho\in\mathscr{P}_{\pi}(\Theta)$,

$$
\log\left[\int_{\Theta}\exp(\mathscr{A}(\theta))\,d\pi(\theta)\right]=\left[\int_{\Theta}\mathscr{A}(\theta)\,d\rho(\theta)-D_{\mathrm{KL}}(\rho,\pi)\right]+D_{\mathrm{KL}}\left(\rho,\rho_{-A,\pi}\right). \quad (2.46)
$$

Note that $D_{\mathrm{KL}}\left(\rho,\rho_{-A,\pi}\right)\geq 0$. It follows from (2.46) that

$$
\begin{aligned}
\log\left[\int_{\Theta}\exp(\mathscr{A}(\theta))\,d\pi(\theta)\right]&=\left[\int_{\Theta}\mathscr{A}(\theta)\,d\rho(\theta)-D_{\mathrm{KL}}(\rho,\pi)\right]+D_{\mathrm{KL}}\left(\rho,\rho_{-A,\pi}\right)\\
&\geq\left[\int_{\Theta}\mathscr{A}(\theta)\,d\rho(\theta)-D_{\mathrm{KL}}(\rho,\pi)\right].
\end{aligned}
$$

This implies that

$$
\int_{\Theta}\mathscr{A}(\theta)\,d\rho(\theta)\leq D_{\mathrm{KL}}(\rho,\pi)+\log\left[\int_{\Theta}\exp(\mathscr{A}(\theta))\,d\pi(\theta)\right].
$$

■

2.A.2 Proofs for Section 2.3.1

Proof of Theorem 2.3.1. Let $A(\theta)=\lambda D[R(\theta),R_n(\theta)]$ and $\lambda\in I$. (2.25) and Fubini's theorem imply that

$$
\int_{\Theta}\exp(\lambda D[R(\theta),R_n(\theta)])\,d\pi(\theta)<\infty
$$

holds almost surely. Therefore, by Corollary 2.2.1 (b), the event

$$\begin{aligned}&\left\{\int_\Theta \lambda D\left[R(\theta), R_n(\theta)\right] d\rho(\theta)\right.\\&\left.\le \log\left[\int_\Theta \exp\left(\lambda D\left[R(\theta), R_n(\theta)\right]\right) d\pi(\theta)\right] + D_{\mathrm{KL}}(\rho,\pi) \text{ for all } \rho \in \mathscr{P}_\pi(\Theta) \text{ simultaneously}\right\}\end{aligned} \tag{2.47}$$

occurs with probability one. Applying Jensen's inequality to the object on the left-hand side of the inequality in this event, we obtain that

$$\begin{aligned}&\Pr\left\{\lambda D\left[R\left(a_{G,\rho}\right), R_n\left(a_{G,\rho}\right)\right]\right.\\&\left.\le \log\left[\int_\Theta \exp\left(\lambda D\left[R(\theta), R_n(\theta)\right]\right) d\pi(\theta)\right] + D_{\mathrm{KL}}(\rho,\pi) \text{ for all } \rho \in \mathscr{P}_\pi(\Theta) \text{ simultaneously}\right\}\\&= 1.\end{aligned} \tag{2.48}$$

Next, we establish a high-probability bound for $\log\left[\int_\Theta \exp\left(\lambda D\left[R(\theta), R_n(\theta)\right]\right) d\pi(\theta)\right]$ using the Markov inequality: for any constant C,

$$\begin{aligned}&\Pr\left\{\log\left[\int_\Theta \exp\left(\lambda D\left[R(\theta), R_n(\theta)\right]\right) d\pi(\theta)\right] > C\right\}\\&\le \Pr\left\{\left[\int_\Theta \exp\left(\lambda D\left[R(\theta), R_n(\theta)\right]\right) d\pi(\theta)\right] > \exp C\right\}\\&\le \frac{E\left[\int_\Theta \exp\left(\lambda D\left[R(\theta), R_n(\theta)\right]\right) d\pi(\theta)\right]}{\exp C}\\&= \frac{\int_\Theta E \exp\left(\lambda D\left[R(\theta), R_n(\theta)\right]\right) d\pi(\theta)}{\exp C} \le \exp\left(f(\lambda, n) - C\right).\end{aligned}$$

where the equality follows from Fubini's theorem and the last inequality follows from (2.25).

Solving the equation $\exp\left(f(\lambda, n) - C\right) = \varepsilon$ for C, we find

$$C = f(\lambda, n) + \log\frac{1}{\varepsilon}.$$

So

$$\Pr\left\{\log\left[\int_{\Theta}\exp\left(\lambda D\left[R\left(\theta\right),R_n\left(\theta\right)\right]\right)d\pi\left(\theta\right)\right]\leq f\left(\lambda,n\right)+\log\frac{1}{\varepsilon}\right\}\geq 1-\varepsilon.$$

Note that the above high probability bound does not involve ρ. Combining this with (2.48), we have

$$\Pr\left\{D\left[R\left(a_{G,\rho}\right),R_n\left(a_{G,\rho}\right)\right]\leq\frac{f\left(\lambda,n\right)+\log\frac{1}{\varepsilon}+D_{\mathrm{KL}}\left(\rho,\pi\right)}{\lambda}\text{ for all }\rho\in\mathscr{P}_{\pi}(\Theta)\text{ simultaneously}\right\}$$
$$\geq 1-\varepsilon. \tag{2.49}$$

■

The proof of Lemma 2.3.1 below will utilize the following two lemmas.

Lemma 2.A.1 *Let X be a random variable with $EX=0$ such that for some constant $K>0$, the MGF of X^2 satisfies*

$$E\exp\left(\lambda^2X^2\right)\leq\exp\left(K^2\lambda^2\right)\text{ for all }\lambda\text{ such that }|\lambda|\leq\frac{1}{K}. \tag{2.50}$$

Then

$$E\exp\left(\lambda X\right)\leq\exp\left(K^2\lambda^2\right)\text{ for all }\lambda\in\mathbb{R}.$$

Proof of Lemma 2.A.1. This follows from the proof of Proposition 2.5.2 in Vershynin (2018), pages 22-23. ■

Lemma 2.A.2 *Let X be any random variable taking values in $[0,1]$ with $EX=\mu$. Denote $\mathbf{X}=(X_1,\ldots,X_n)$ where $X_1,\ldots,X_n$ are iid realizations of X. Let $\mathbf{X}'=(X_1',\ldots,X_n')$ where $X_1',\ldots,X_n'$ are iid realizations of a Bernoulli random variable X' with probability of success μ. If $f:[0,1]^n\to\mathbb{R}$ is convex, then*

$$E\left[f\left(\mathbf{X}\right)\right]\leq E\left[f\left(\mathbf{X}'\right)\right]$$

Proof of Lemma 2.A.2. This lemma is due to Maurer (2004). Another proof with more details is given in Germain et al. (2015); see Lemmas 51 and 52 there. For intuition, we can regard $\mathbf{X}'$ as a mean-preserving spread of $\mathbf{X}$ and $-f$ as the utility function. Then the lemma says that $\mathbf{X}$ is preferred by an expected utility maximizer having concave utility $-f(\cdot)$. ■

Proof of Lemma 2.3.1.

Part (a). Let $(X,Y) \sim P(X,Y)$ and let $\mu_\psi = E\psi(X,Y) < \infty$ where finiteness follows from Assumption 2.2.2 (iv). Recall that under Assumption 2.2.2 (iv), there exists a constant $K_\psi > 0$ such that

$$E\exp\left\{\lambda^2\psi(X,Y)^2\right\} \leq \exp\left(K_\psi^2\lambda^2\right) \text{ for all } \lambda \text{ such that } |\lambda| \leq \frac{1}{K_\psi}. \tag{2.51}$$

Now for either $s \in \{-1,1\}$ and any $\theta \in \Theta$, consider

$$s\left[E\ell(\theta,Y,X) - \ell(\theta,Y,X)\right] = s\left[E\left(\psi(X,Y)1\{Y \neq a(X,\theta)\}\right) - \psi(X,Y)1\{Y \neq a(X,\theta)\}\right]. \tag{2.52}$$

Recall that $\psi(X,Y) > 0$. Using $(a-b)^2 \leq a^2 + b^2$ for $a > 0$ and $b > 0$, we have

$$\begin{aligned}&E\exp\left\{\lambda^2\left(s\left[E\left\{\psi(X,Y)1\{Y \neq a(X,\theta)\}\right\} - \psi(X,Y)1\{Y \neq a(X,\theta)\}\right]\right)^2\right\}\\ &\leq E\exp\left\{\lambda^2\psi(X,Y)^2 + \lambda^2\left(E\left\{\psi(X,Y)\right\}\right)^2\right\}.\end{aligned}$$

Additionally,

$$E\exp\left\{\lambda^2\psi(X,Y)^2 + \lambda^2\mu_\psi^2\right\} \leq \exp\left(\lambda^2\left[K_\psi^2 + \mu_\psi^2\right]\right)$$

for any λ such that $|\lambda| \leq 1/K_\psi$, which follows from (2.51). Seeing as $1/(K_\psi^2+\mu_\psi^2)^{1/2} < 1/K_\psi$, the following condition holds

$$E\exp\left\{\lambda^2\left(s\left[E\ell(\theta,Y,X) - \ell(\theta,Y,X)\right]\right)^2\right\} \leq \exp\left(\lambda^2\left[K_\psi^2 + \mu_\psi^2\right]\right),$$

for all λ such that $|\lambda| \leq 1/(K_\psi^2+\mu_\psi^2)^{1/2}$. As the expression in (2.52) has mean zero, Lemma

2.A.1 yields that

$$E\exp\left\{\lambda\left(s\left[E\ell\left(\theta,Y,X\right)-\ell\left(\theta,Y,X\right)\right]\right)\right\}\leq\exp\left(\left[K_{\psi}^{2}+\mu_{\psi}^{2}\right]\lambda^{2}\right)\text{ for all }\lambda\in\mathbb{R}.\qquad(2.53)$$

Applying (2.53),

$$\begin{aligned}E\exp\left\{\lambda D\left[R\left(\theta\right),R_{n}\left(\theta\right)\right]\right\}&=E\exp\left\{\lambda s\left[R\left(\theta\right)-R_{n}\left(\theta\right)\right]\right\}\\&=E\exp\left\{\sum_{i=1}^{n}\left[\frac{\lambda}{n}\left(s\left[E\ell\left(\theta,Y_{i},X_{i}\right)-\ell\left(\theta,Y_{i},X_{i}\right)\right]\right)\right]\right\}\\&=\prod_{i=1}^{n}E\exp\left\{\frac{\lambda}{n}\left(s\left[E\ell\left(\theta,Y_{i},X_{i}\right)-\ell\left(\theta,Y_{i},X_{i}\right)\right]\right)\right\}\\&\leq\prod_{i=1}^{n}\exp\left(\frac{\lambda^{2}\left[K_{\psi}^{2}+\mu_{\psi}^{2}\right]}{n^{2}}\right)=\exp\left(\frac{\lambda^{2}\left[K_{\psi}^{2}+\mu_{\psi}^{2}\right]}{n}\right).\end{aligned}$$

Taking an integral with respect to π yields

$$\int_{\Theta}E\exp\left\{\lambda D\left[R\left(\theta\right),R_{n}\left(\theta\right)\right]\right\}d\pi(\theta)\leq\exp\left(\frac{\lambda^{2}\left[K_{\psi}^{2}+\mu_{\psi}^{2}\right]}{n}\right),$$

implying the first expression for $f(\lambda,n)$.

To derive the second expression for $f(\lambda,n)$ in the case that the utility function is bounded, note that $\psi(x,y)=U(1,1,x)-U(-1,1,x)$ when $y=1$ and $\psi(x,y)=U(-1,-1,x)-U(1,-1,x)$ when $y=-1$. When

$$U_{\max}=\sup_{a,y,x}|U(a,y,x)|<\infty,$$

it follows that $0\leq\ell(\theta,y,x)=\psi(x,y)1\{y\neq\operatorname{sign}[m(x,\theta)-c(x)]\}<2U_{\max}$ under Assumption 2.2.1. By Hoeffding's lemma (see, for example, Massart and Picard (2007), page 21), with

$s = -1$, for any $\theta \in \Theta$ we have

$$\begin{aligned} E\exp\left(\lambda\left[R_n(\theta) - R(\theta)\right]\right) &= E\exp\left(\frac{\lambda}{n}\sum_{i=1}^{n}\left[\ell(\theta, Y_i, X_i) - E\ell(\theta, Y_i, X_i)\right]\right) \\ &= \prod_{i=1}^{n} E\exp\left\{\frac{\lambda}{n}\left[\ell(\theta, Y_i, X_i) - E\ell(\theta, Y_i, X_i)\right]\right\} \\ &\le \prod_{i=1}^{n}\exp\left(\frac{\lambda^2 U_{\max}^2}{2n^2}\right) = \exp\left(\frac{\lambda^2 U_{\max}^2}{2n}\right). \end{aligned} \tag{2.54}$$

Nearly identical steps in the $s = 1$ case, now with Hoeffding's lemma applied to $-\ell(\theta, Y_i, X_i)$, $i = 1, \ldots, n$, produce that

$$E\exp\left(\lambda\left[R(\theta) - R_n(\theta)\right]\right) \le \exp\left(\frac{\lambda^2 U_{\max}^2}{2n}\right) \tag{2.55}$$

Integrating with respect to π, (2.54) and (2.55) yield that

$$\int_{\Theta} E\exp\left(\lambda s\left[R(\theta) - R_n(\theta)\right]\right) d\pi(\theta) \le \exp\left(\frac{\lambda^2 U_{\max}^2}{2n}\right),\ s \in \{-1, 1\}.$$

This demonstrates that (2.25) holds with $f(\lambda, n) = \frac{\lambda^2 U_{\max}^2}{2n}$ in the bounded utility setting.

Part (b). Again note that when the utility function is bounded by $U_{\max}$ we have $0 \le \ell(\theta, y, x) < 2U_{\max}$ under Assumption 2.2.1. Therefore $\ell(\theta, y, x)/(2U_{\max}) \in [0, 1]$. Set

$$\mathbf{X} = \left(\frac{\ell(\theta, Y_1, X_1)}{2U_{\max}}, \ldots, \frac{\ell(\theta, Y_n, X_n)}{2U_{\max}}\right),$$

and note that for any $\theta \in \Theta$,

$$\begin{aligned} \exp\{\lambda D(R(\theta), R_n(\theta))\} &= \exp\left[\lambda\mathscr{F}(R(\theta)) - 2U_{\max}\lambda \cdot \frac{R_n(\theta)}{2U_{\max}}\right] \\ &= \exp\left\{\lambda\mathscr{F}(R(\theta)) - \frac{2U_{\max}\lambda}{n}\sum_{i=1}^{n}\frac{\ell(\theta, Y_i, X_i)}{2U_{\max}}\right\} \end{aligned} \tag{2.56}$$

is a convex mapping of **X**. By (Maurer's) Lemma 2.A.2,

$$E\exp\left\{\lambda\mathscr{F}\left(R(\theta)\right)-\frac{2U_{\max}\lambda}{n}\sum_{i=1}^{n}\frac{\ell(\theta,Y_i,X_i)}{2U_{\max}}\right\}\leq E\exp\left\{\lambda\mathscr{F}\left(R(\theta)\right)-\frac{2U_{\max}\lambda}{n}\sum_{i=1}^{n}X_i'\right\}, \tag{2.57}$$

where $X_1',\ldots,X_n'$ are iid Bernoulli random variables with success probability $R(\theta)/\left(2U_{\max}\right)\in[0,1]$. From here we can continue as in the proof of Corollary 2.2 in Germain et al. (2009). We have for any $\theta\in\Theta$,

$$\begin{aligned}
&E\exp\left\{\lambda\mathscr{F}\left(R(\theta)\right)-\frac{2U_{\max}\lambda}{n}\sum_{i=1}^{n}X_i'\right\}\\
&=\exp\left\{\lambda\mathscr{F}\left(R(\theta)\right)\right\}E\exp\left\{-\frac{2U_{\max}\lambda}{n}\sum_{i=1}^{n}X_i'\right\}\\
&=\exp\left\{\lambda\mathscr{F}\left(R(\theta)\right)\right\}\sum_{k=1}^{n}\Pr\left(\sum_{i=1}^{n}X_i'=k\right)\exp\left(-\frac{2U_{\max}\lambda}{n}k\right)\\
&=\exp\left\{\lambda\mathscr{F}\left(R(\theta)\right)\right\}\sum_{k=1}^{n}\binom{n}{k}\left(\frac{R(\theta)}{2U_{\max}}\right)^{k}\left(1-\frac{R(\theta)}{2U_{\max}}\right)^{n-k}\left[\exp\left(-\frac{2U_{\max}\lambda}{n}\right)\right]^{k}\\
&=\exp\left\{\lambda\mathscr{F}\left(R(\theta)\right)\right\}\left[\left(\frac{R(\theta)}{2U_{\max}}\right)\exp\left(-\frac{2U_{\max}\lambda}{n}\right)+\left(1-\frac{R(\theta)}{2U_{\max}}\right)\right]^{n}\\
&=\exp\left\{\lambda\mathscr{F}\left(R(\theta)\right)\right\}\left\{1-\left(\frac{R(\theta)}{2U_{\max}}\right)\left[1-\exp\left(-\frac{2U_{\max}\lambda}{n}\right)\right]\right\}^{n},
\end{aligned}$$

where the second to last equality is from the binomial theorem. Now, noting that

$$\exp\left\{\lambda\mathscr{F}\left(R(\theta)\right)\right\}=\left\{1-\left(\frac{R(\theta)}{2U_{\max}}\right)\left[1-\exp\left(-\frac{2U_{\max}\lambda}{n}\right)\right]\right\}^{-n},$$

we have

$$E\exp\left\{\lambda\mathscr{F}\left(R(\theta)\right)-\frac{2U_{\max}\lambda}{n}\sum_{i=1}^{n}X_i'\right\}=1. \tag{2.58}$$

Combining equations (2.56), (2.57), and (2.58), we have

$$E\exp\left\{\lambda\left[\mathscr{F}\left(R(\theta)\right)-R_n\left(\theta\right)\right]\right\}\leq 1, \tag{2.59}$$

and so equation (2.25) holds with $f(\lambda, n) = 0$.

Part (c). Let $\theta \in \Theta$. If $R(\theta) - \lambda U_{\max}^2/(2n) \geq \mathscr{F}(R(\theta))$, which is not random, then clearly

$$D(R(\theta), R_n(\theta)) = R(\theta) - \lambda U_{\max}^2/(2n) - R_n(\theta).$$

In this case,

$$\begin{aligned} E\exp\left(\lambda D[R(\theta), R_n(\theta)]\right) &= E\exp\left(\lambda\left[R(\theta) - \frac{\lambda U_{\max}^2}{2n} - R_n(\theta)\right]\right) \\ &= \exp\left(-\frac{\lambda^2 U_{\max}^2}{2n}\right) E\exp\left(\lambda\left[R(\theta) - R_n(\theta)\right]\right) \\ &\leq \exp\left(-\frac{\lambda^2 U_{\max}^2}{2n}\right)\exp\left(\frac{\lambda^2 U_{\max}^2}{2n}\right) = 1 \qquad (2.60) \end{aligned}$$

where the inequality follows from (2.55). Alternatively, in the case that $R(\theta) - \lambda U_{\max}^2/(2n) < \mathscr{F}(R(\theta))$, we have

$$D[R(\theta), R_n(\theta)] = \mathscr{F}(R(\theta)) - R_n(\theta).$$

Then, by (2.59),

$$E\exp\left(\lambda D[R(\theta), R_n(\theta)]\right) = E\exp\left(\lambda\left[\mathscr{F}\left(R(\theta)\right) - R_n(\theta)\right]\right) \leq 1. \qquad (2.61)$$

Integrating over Θ with respect to π, it follows from (2.60) and (2.61) that

$$\int_{\Theta} E\exp\left(\lambda D\left[R(\theta), R_n(\theta)\right]\right) d\pi(\theta) \leq 1,$$

so condition (2.25) holds with $f(\lambda, n) = 0$. ■

Proof of Theorem 2.3.2.

Part (a) follows directly from Theorem 2.3.1 and Lemma 2.3.1 (a) with D as in Lemma 2.3.1 (a).

Part (b). Let D be as specified in Lemma 2.3.1 (b). It is straightforward to verifty that D is convex. Note that

$$D(r_1, r_2) = \mathscr{F}_{\lambda,n}(r_1) - r_2 \leq \mathfrak{d}$$

for any $\mathfrak{d} \in \mathbb{R}$ if and only if

$$1 - \left(\frac{r_1}{2U_{\max}}\right)\left[1 - \exp\left(-\frac{2U_{\max}\lambda}{n}\right)\right] \geq \exp\left[-\frac{\lambda}{n}(r_2 + \mathfrak{d})\right].$$

The latter is equivalent to

$$\begin{aligned} r_1 &\leq \frac{2U_{\max}}{1 - \exp(-2U_{\max}\lambda/n)}\left\{1 - \exp\left[-\frac{\lambda}{n}(r_2 + \mathfrak{d})\right]\right\} \\ &:= \mathscr{F}^{-1}_{\lambda,n}(r_2 + \mathfrak{d}). \end{aligned}$$

Setting $r_1 = \int_\Theta R(\theta)\, d\rho(\theta)$, $r_2 = \int_\Theta R_n(\theta)\, d\rho(\theta)$ and $\mathfrak{d} = \frac{1}{\lambda}\left[\log\frac{1}{\varepsilon} + D_{\mathrm{KL}}(\rho,\pi)\right]$ and using Theorem 2.3.1 and Lemma 2.3.1 (b) yields the desired result.

Part (c). Now let D be as specified in Lemma 2.3.1 (c). That D is convex follows from the convexity of D specified in part (a) plus a constant, the convexity of D specified in part (b), and the fact that the maximum of two convex functions is convex. Theorem 2.3.1 combined with Lemma 2.3.1 (c) yields that

$$\begin{aligned} \Pr\Bigg\{ \max\left[\int_\Theta R(\theta)\, d\rho(\theta) - \frac{\lambda U_{\max}^2}{2n} - \int_\Theta R_n(\theta)\, d\rho(\theta),\ \mathscr{F}\left(\int_\Theta R(\theta)\, d\rho(\theta)\right) - \int_\Theta R_n(\theta)\, d\rho(\theta)\right] \\ \leq \frac{D_{\mathrm{KL}}(\rho,\pi) + \log\frac{1}{\varepsilon}}{\lambda}\ \mathit{for\ all}\ \rho \in \mathscr{P}_\pi(\Theta)\ \mathit{simultaneously}\Bigg\} \geq 1 - \varepsilon \end{aligned} \tag{2.62}$$

Now, observe that

$$\begin{aligned} \max\left[\int_\Theta R(\theta)\, d\rho(\theta) - \frac{\lambda U_{\max}^2}{2n} - \int_\Theta R_n(\theta)\, d\rho(\theta),\ \mathscr{F}\left(\int_\Theta R(\theta)\, d\rho(\theta)\right) - \int_\Theta R_n(\theta)\, d\rho(\theta)\right] \\ \leq \frac{D_{\mathrm{KL}}(\rho,\pi) + \log\frac{1}{\varepsilon}}{\lambda} \end{aligned}$$

holds if and only if

$$\int_\Theta R(\theta)\,d\rho(\theta) \leq \int_\Theta R_n(\theta) d\rho(\theta) + \frac{1}{\lambda}\left[\frac{\lambda^2 U_{\max}^2}{2n} + D_{\mathrm{KL}}(\rho,\pi) + \log\frac{1}{\varepsilon}\right] = U_{\lambda,\pi,\rho}(\varepsilon),$$

and

$$\int_\Theta R(\theta)\,d\rho(\theta) \leq \mathscr{F}_{n,\lambda}^{-1}\left(\int_\Theta R_n(\theta)\,d\rho(\theta) + \frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi) + \frac{1}{\lambda}\log\frac{1}{\varepsilon}\right) = U_{\lambda,\pi,\rho}^{\mathscr{F}}(\varepsilon)$$

hold simultaneously. Additionally, the two inequalities directly above hold simultaneously if and only if

$$\int_\Theta R(\theta)\,d\rho(\theta) \leq \min\left\{U_{\lambda,\pi,\rho}(\varepsilon), U_{\lambda,\pi,\rho}^{\mathscr{F}}(\varepsilon)\right\}.$$

Therefore,

$$\begin{aligned}
&\left\{\max\left[\int_\Theta R(\theta)\,d\rho(\theta) - \frac{\lambda U_{\max}^2}{2n} - \int_\Theta R_n(\theta)\,d\rho(\theta),\ \mathscr{F}\left(\int_\Theta R(\theta)\,d\rho(\theta)\right) - \int_\Theta R_n(\theta)\,d\rho(\theta)\right]\right.\\
&\qquad\qquad \left.\leq \frac{D_{\mathrm{KL}}(\rho,\pi) + \log\frac{1}{\varepsilon}}{\lambda}\ \mathit{for\ all}\ \rho \in \mathscr{P}_\pi(\Theta)\ \mathit{simultaneously}\right\} \qquad (2.63)\\
&= \left\{\int_\Theta R(\theta)\,d\rho(\theta) \leq \min\left\{U_{\lambda,\pi,\rho}(\varepsilon), U_{\lambda,\pi,\rho}^{\mathscr{F}}(\varepsilon)\right\}\ \mathit{for\ all}\ \rho \in \mathscr{P}_\pi(\Theta)\ \mathit{simultaneously}\right\}.
\end{aligned}$$

Combined, (2.62) and (2.63) imply the result of Theorem 2.3.2 (c). ■

Proof of Theorem 2.3.3. Part (a). This part follows directly from Theorem 2.3.2 with $s = 1$ and $\rho = \hat{\rho}_\lambda$.

Part (b). Define the events $\mathscr{E}_1$ and $\mathscr{E}_2$:

$$\mathscr{E}_1 = \left\{\int_\Theta R(\theta)\,d\hat{\rho}_\lambda(\theta) \leq \int_\Theta R_n(\theta)\,d\hat{\rho}_\lambda(\theta) + \frac{1}{\lambda}\left[D_{\mathrm{KL}}(\hat{\rho}_\lambda,\pi) + \frac{\lambda^2\left(K_\psi^2 + \mu_\psi^2\right)}{n} + \log\frac{2}{\varepsilon}\right]\right\},$$

$$\mathscr{E}_2 = \left\{\int_\Theta R_n(\theta)\,d\hat{\rho}_\lambda(\theta) \leq \int_\Theta R(\theta)\,d\hat{\rho}_\lambda(\theta) + \frac{1}{\lambda}\left[D_{\mathrm{KL}}(\hat{\rho}_\lambda,\pi) + \frac{\lambda^2\left(K_\psi^2 + \mu_\psi^2\right)}{n} + \log\frac{2}{\varepsilon}\right]\right\}.$$

Then

$$\Pr\left(\mathscr{E}_1\right) \geq 1-\frac{\varepsilon}{2} \text{ and } \Pr\left(\mathscr{E}_2\right) \geq 1-\frac{\varepsilon}{2}.$$

So

$$\begin{aligned}\Pr\left(\mathscr{E}_1 \cap \mathscr{E}_2\right) = 1-\Pr\left(\mathscr{E}_1^c \cup \mathscr{E}_2^c\right) &\geq 1-\Pr\left(\mathscr{E}_1^c\right)-\Pr\left(\mathscr{E}_2^c\right) \\ &\geq 1-\frac{\varepsilon}{2}-\frac{\varepsilon}{2}=1-\varepsilon.\end{aligned}$$

But, the event given in Part (b) is just $\mathscr{E}_1 \cap \mathscr{E}_2$. Hence, the inequality in Part (b) holds with probability at least $1-\varepsilon$.

Part (c). By the definition of $\hat{\rho}_\lambda$, we have

$$\int_\Theta R_n(\theta)\, d\hat{\rho}_\lambda(\theta)+\frac{1}{\lambda} D_{\mathrm{KL}}\left(\hat{\rho}_\lambda, \pi\right) \leq \int_\Theta R_n(\theta)\, d\rho(\theta)+\frac{1}{\lambda} D_{\mathrm{KL}}(\rho, \pi)$$

for all $\rho \in \mathscr{P}_\pi(\Theta)$ simultaneously. Hence, by part (a), with probability at least $1-\varepsilon/2$:

$$\int_\Theta R(\theta)\, d\hat{\rho}_\lambda(\theta) \leq \int_\Theta R_n(\theta)\, d\rho(\theta)+\frac{1}{\lambda}\left[D_{\mathrm{KL}}(\rho, \pi)+\frac{\lambda^2\left(K_\psi^2+\mu_\psi^2\right)}{n}+\log \frac{2}{\varepsilon}\right]$$

for all $\rho \in \mathscr{P}_\pi(\Theta)$ simultaneously. Using Theorem 2.3.2 (a) now with $s=-1$, we have, with probability at least $1-\varepsilon/2$,

$$\int_\Theta R_n(\theta)\, d\rho(\theta) \leq \int_\Theta R(\theta)\, d\rho(\theta)+\frac{1}{\lambda}\left[D_{\mathrm{KL}}(\rho, \pi)+\frac{\lambda^2\left(K_\psi^2+\mu_\psi^2\right)}{n}+\log \frac{2}{\varepsilon}\right].$$

Therefore, with probability at least $1-\varepsilon$,

$$\int_\Theta R(\theta)\, d\hat{\rho}_\lambda(\theta) \leq \int_\Theta R(\theta)\, d\rho(\theta)+\frac{2}{\lambda} D_{\mathrm{KL}}(\rho, \pi)+\frac{2}{\lambda}\left[\frac{\lambda^2\left(K_\psi^2+\mu_\psi^2\right)}{n}+\log \frac{2}{\varepsilon}\right] \tag{2.64}$$

for all $\rho \in \mathscr{P}_{\pi}(\Theta)$ simultaneously. Hence, with probability at least $1-\varepsilon$,

$$\int_{\Theta} R(\theta)\, d\hat{\rho}_{\lambda}(\theta) \leq \sup_{\rho \in \mathscr{P}_{\pi}(\Theta)} \left[\int_{\Theta} R(\theta)\, d\rho(\theta) + \frac{2}{\lambda} D_{\mathrm{KL}}(\rho, \pi) \right] + \frac{2}{\lambda} \left[\frac{\lambda^2 \left(K_{\psi}^2 + \mu_{\psi}^2\right)}{n} + \log \frac{2}{\varepsilon} \right].$$

In the case where $U_{\max} < \infty$, we can follow the same steps based off Theorem 2.3.2 but with $(K_{\psi}^2 + \mu_{\psi}^2)$ replaced by $U_{\max}^2/2$.

Part (d). This follows directly from Theorem 2.3.2(c) with $\rho = \hat{\rho}_{\lambda}$. ■

Proof of Theorem 2.3.4.

Part (a). Let $s \in \{0,1\}$. For any $\rho \in \mathscr{P}(\Theta)_{\pi}$, including sample dependent ρ, let

$$B_n(\lambda_1, \lambda_2, z; \rho, \pi) = \frac{1}{\lambda_1} \left[\frac{\lambda_2^2 U_{\max}^2}{2n} + \log z + D_{\mathrm{KL}}(\rho, \pi) \right],$$

and define the event

$$\mathscr{E}_n(\lambda_1, \lambda_2, z; \rho, \pi) = \left\{ \int_{\Theta} s\left[R(\theta) - R_n(\theta)\right] d\rho(\theta) > B_n(\lambda_1, \lambda_2, z; \rho, \pi) \right\}.$$

Note that by Theorem 2.3.2(a), $\Pr(\mathscr{E}_n(\lambda, \lambda, 1/\varepsilon; \rho, \pi)) \leq \varepsilon$ for any $\lambda > 0$. Additionally, holding the other arguments constant, $B_n(\lambda_1, \lambda_2, z; \rho, \pi)$ is decreasing in λ_1, increasing in λ_2, and increasing in z. Hence

$$\begin{aligned}
\mathscr{E}_n(\lambda_1, \lambda_2, z; \rho, \pi) &\subseteq \mathscr{E}_n\left(\tilde{\lambda}_1, \lambda_2, z; \rho, \pi\right) \text{ for } \tilde{\lambda}_1 \geq \lambda_1, \\
\mathscr{E}_n(\lambda_1, \lambda_2, z; \rho, \pi) &\subseteq \mathscr{E}_n\left(\lambda_1, \tilde{\lambda}_2, z; \rho, \pi\right) \text{ for } \tilde{\lambda}_2 \leq \lambda_2, \\
\mathscr{E}_n(\lambda_1, \lambda_2, z; \rho, \pi) &\subseteq \mathscr{E}_n(\lambda_1, \lambda_2, \tilde{z}; \rho, \pi) \text{ for } \tilde{z} \leq z.
\end{aligned}$$

Now, fix $\alpha > 1$. With some abuse of notation, the event of interest in Part (a) is the complement of the event $\mathscr{E}_n(\alpha;\rho,\pi)$ defined by

$$\mathscr{E}_n(\alpha;\rho,\pi) := \left\{ \int_{\Theta} s\left[R(\theta) - R_n(\theta)\right] d\rho(\theta) > \inf_{\lambda>1} \left\{ B_n\left(\frac{\lambda}{\alpha}, \lambda, \frac{1}{\varepsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2; \rho, \pi\right) \right\} \right\}.$$

Note that

$$\mathscr{E}_n(\alpha;\rho,\pi) = \bigcup_{\lambda>1} \mathscr{E}_n\left(\frac{\lambda}{\alpha}, \lambda, \frac{1}{\varepsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2; \rho, \pi\right).$$

But

$$\bigcup_{\lambda>1} \mathscr{E}_n\left(\frac{\lambda}{\alpha}, \lambda, \frac{1}{\varepsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2; \rho, \pi\right) \subseteq \bigcup_{k=0}^{\infty} \bigcup_{\lambda\in(\alpha^k,\alpha^{k+1}]} \mathscr{E}_n\left(\frac{\lambda}{\alpha}, \lambda, \frac{1}{\varepsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2; \rho, \pi\right),$$

and for all $\lambda \in (\alpha^k, \alpha^{k+1}]$ it holds that

$$\mathscr{E}_n\left(\frac{\lambda}{\alpha}, \lambda, \frac{1}{\varepsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2; \rho, \pi\right) \subseteq \mathscr{E}_n\left(\frac{\alpha^{k+1}}{\alpha}, \alpha^k, \frac{1}{\varepsilon}\left(\frac{\log\left(\alpha^2\cdot\alpha^k\right)}{\log\alpha}\right)^2; \rho, \pi\right).$$

Hence

$$\bigcup_{\lambda\in(\alpha^k,\alpha^{k+1}]} \mathscr{E}_n\left(\frac{\lambda}{\alpha}, \lambda, \frac{1}{\varepsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2; \rho, \pi\right) \subseteq \mathscr{E}_n\left(\frac{\alpha^{k+1}}{\alpha}, \alpha^k, \frac{1}{\varepsilon}\left(\frac{\log\left(\alpha^2\cdot\alpha^k\right)}{\log\alpha}\right)^2; \rho, \pi\right),$$

and

$$\begin{aligned}
\Pr\left(\mathscr{E}_n(\alpha;\rho,\pi)\right) &\leq \sum_{k=0}^{\infty} \Pr\left[\bigcup_{\lambda\in(\alpha^k,\alpha^{k+1}]} \mathscr{E}_n\left(\frac{\lambda}{\alpha}, \lambda, \frac{1}{\varepsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2; \rho, \pi\right)\right] \\
&\leq \sum_{k=0}^{\infty} \Pr\left[\mathscr{E}_n\left(\frac{\alpha^{k+1}}{\alpha}, \alpha^k, \frac{1}{\varepsilon}\left(\frac{\log\left(\alpha^2\cdot\alpha^k\right)}{\log\alpha}\right)^2; \rho, \pi\right)\right] \\
&= \sum_{k=0}^{\infty} \Pr\left[\mathscr{E}_n\left(\alpha^k, \alpha^k, \frac{(k+2)^2}{\varepsilon}; \rho, \pi\right)\right] \\
&\leq \sum_{k=0}^{\infty} \frac{\varepsilon}{(k+2)^2} = \left(\frac{1}{6}\pi^2 - 1\right)\varepsilon < \varepsilon,
\end{aligned}$$

where last inequality follows from Theorem 2.3.2(a). Therefore,

$$\Pr\left(\mathscr{E}_n\left(\alpha;\rho,\pi\right)^c\right) \geq 1-\varepsilon$$

This is the statement for Part (a).

Part (b). Applying Part (a) with $\rho = \hat{\rho}_{\tilde{\lambda}}$, the following event holds with probability probability $1-\varepsilon$

$$\int_\Theta s\left[R(\theta)-R_n(\theta)\right] d\hat{\rho}_{\tilde{\lambda}}(\theta) \leq \inf_{\lambda>1}\left\{\frac{\alpha}{\lambda}\left[\frac{\lambda^2 U_{\max}^2}{2n}+\log\frac{1}{\varepsilon}+D_{\mathrm{KL}}(\hat{\rho}_{\tilde{\lambda}},\pi)+2\log\frac{\log\left(\alpha^2\lambda\right)}{\log\alpha}\right]\right\}.$$

Then, Part (b) follows from the above and the observation that, for $\tilde{\lambda} > 1$, it holds that

$$\begin{aligned}\inf_{\lambda>1}&\left\{\frac{\alpha}{\lambda}\left[\frac{\lambda^2 U_{\max}^2}{2n}+\log\frac{1}{\varepsilon}+D_{\mathrm{KL}}(\hat{\rho}_{\tilde{\lambda}},\pi)+2\log\frac{\log\left(\alpha^2\lambda\right)}{\log\alpha}\right]\right\}\\ &\leq\frac{\alpha}{\tilde{\lambda}}\left[\frac{\tilde{\lambda}^2 U_{\max}^2}{2n}+\log\frac{1}{\varepsilon}+D_{\mathrm{KL}}(\hat{\rho}_{\tilde{\lambda}},\pi)+2\log\frac{\log\left(\alpha^2\tilde{\lambda}\right)}{\log\alpha}\right].\end{aligned}$$

Part (c). We proceed similarly to part (a). For any $\rho \in \mathscr{P}_\pi(\Theta)$ that may be sample dependent, define

$$\overline{B}_n\left(\lambda_1,\lambda_2,z;\rho,\pi\right)=\int_\Theta R_n(\theta)d\rho(\theta)+\frac{1}{\lambda_1}\left[\frac{\lambda_2^2 U_{\max}^2}{2n}+\log z+D_{\mathrm{KL}}(\rho,\pi)\right],$$

and

$$\begin{aligned}&\overline{B}_n^{\mathscr{F}}\left(\lambda_1,\lambda_2,z;\rho,\pi\right)\\ &\quad=\frac{2U_{\max}}{1-\exp\left(-\frac{2\lambda_1 U_{\max}}{n}\right)}\left\{1-\exp\left[-\frac{\lambda_2}{n}\int_\Theta R_n d\rho(\theta)-\frac{1}{n}\log z-\frac{1}{n}D_{\mathrm{KL}}\left(\rho,\pi\right)\right]\right\}.\end{aligned}$$

Note that

$$\overline{B}_n\left(\frac{\lambda}{\alpha},\lambda,\frac{1}{\varepsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right);\rho,\pi\right)=\overline{U}_{\lambda,\pi,\rho,\alpha}(\varepsilon),$$

and

$$\begin{aligned}
&\overline{B}_n^{\mathscr{F}}\left(\frac{\lambda}{\alpha},\lambda,\frac{1}{\varepsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2;\rho,\pi\right)\\
&=\frac{2U_{\max}}{1-\exp\left(-\frac{2\lambda U_{\max}}{\alpha n}\right)}\left\{1-\exp\left[-\frac{\lambda}{n}\int_\Theta R_n d\rho(\theta)-\frac{1}{n}\log\frac{1}{\varepsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2-\frac{1}{n}D_{\mathrm{KL}}(\rho,\pi)\right]\right\}\\
&=\frac{2U_{\max}}{1-\exp\left(-\frac{2\lambda U_{\max}}{\alpha n}\right)}\left\{1-\exp\left[-\frac{\lambda}{n}\left(\int_\Theta R_n d\rho(\theta)+\frac{1}{\lambda}\log\frac{1}{\varepsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2+\frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right)\right]\right\}\\
&=\mathscr{F}_{n,\lambda,\alpha}^{-1}\left(\int_\Theta R_n d\rho(\theta)+\frac{1}{\lambda}\log\frac{1}{\varepsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2+\frac{1}{\lambda}D_{\mathrm{KL}}(\rho,\pi)\right)=\overline{U}_{\lambda,\pi,\rho,\alpha}^{\mathscr{F}}(\varepsilon).
\end{aligned}$$

Now, holding the other arguments constant, notice that

$$\min[\overline{B}_n(\lambda_1,\lambda_2,z;\rho,\pi),\overline{B}_n^{\mathscr{F}}(\lambda_1,\lambda_2,z;\rho,\pi)]$$

is decreasing in λ_1, increasing in λ_2, and increasing in z as $\overline{B}_n(\lambda_1,\lambda_2,z;\rho,\pi)$, $\overline{B}_n^{\mathscr{F}}(\lambda_1,\lambda_2,z;\rho,\pi)$ both have these properties.

With some abuse of notation, define the two events:

$$\overline{\mathscr{E}}_n(\lambda_1,\lambda_2,z;\rho,\pi)=\left\{\int_\Theta R(\theta)d\rho(\theta)>\min\left[\overline{B}_n(\lambda_1,\lambda_2,z;\rho,\pi),\overline{B}_n^{\mathscr{F}}(\lambda_1,\lambda_2,z;\rho,\pi)\right]\right\},$$

$$\overline{\mathscr{E}}_n(\alpha;\rho,\pi)=\left\{\int_\Theta R(\theta)d\rho(\theta)>\inf_{\lambda>1}\left\{\min\left[\overline{B}_n(\lambda_1,\lambda_2,z;\rho,\pi),\overline{B}_n^{\mathscr{F}}(\lambda_1,\lambda_2,z;\rho,\pi)\right]\right\}\right\}$$

and notice that by Theorem 2.3.2 (c), $\Pr(\overline{\mathscr{E}}_n(\lambda,\lambda,1/\varepsilon;\rho,\pi))\leq\varepsilon$ for any $\lambda>0$.

The event of interest in Part (c) is the complement of $\overline{\mathscr{E}}(\alpha;\rho,\pi)$. Now, we have

$$\begin{aligned}\overline{\mathscr{E}}_n(\alpha;\rho,\pi) &= \bigcup_{\lambda>1}\overline{\mathscr{E}}_n\left(\frac{\lambda}{\alpha},\lambda,\frac{1}{\varepsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2;\rho,\pi\right)\\ &\subseteq \bigcup_{k=0}^{\infty}\bigcup_{\lambda\in(\alpha^k,\alpha^{k+1}]}\overline{\mathscr{E}}_n\left(\frac{\lambda}{\alpha},\lambda,\frac{1}{\varepsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2;\rho,\pi\right).\end{aligned}$$

Hence, following arguments similar to those in part (a),

$$\begin{aligned}\Pr\left(\overline{\mathscr{E}}_n(\alpha;\rho,\pi)\right) &\leq \sum_{k=0}^{\infty}\Pr\left[\bigcup_{\lambda\in(\alpha^k,\alpha^{k+1}]}\overline{\mathscr{E}}_n\left(\frac{\lambda}{\alpha},\lambda,\frac{1}{\varepsilon}\left(\frac{\log\alpha^2\lambda}{\log\alpha}\right)^2;\rho,\pi\right)\right]\\ &\leq \sum_{k=0}^{\infty}\Pr\left[\overline{\mathscr{E}}_n\left(\frac{\alpha^{k+1}}{\alpha},\alpha^k,\frac{1}{\varepsilon}\left(\frac{\log\left(\alpha^2\cdot\alpha^k\right)}{\log\alpha}\right)^2;\rho,\pi\right)\right]\\ &= \sum_{k=0}^{\infty}\Pr\left[\overline{\mathscr{E}}_n\left(\alpha^k,\alpha^k,\frac{(k+2)^2}{\varepsilon};\rho,\pi\right)\right]\\ &< \varepsilon,\end{aligned}$$

It follows that $\Pr\left(\tilde{\mathscr{E}}_n(\alpha;\rho,\pi)^c\right)\geq 1-\varepsilon$, which is the statement of interest for part (c).

Part (d) follows from Part (c) via steps parallel to those in the proof of Part (b). ■

Proof of Lemma 2.3.2. We will show that for all $\theta\in\Theta$,

$$E\exp\left\{\lambda D\left(\frac{R(\theta)}{2U_{\max}},\frac{R_n(\theta)}{2U_{\max}}\right)\right\} = E\exp\left\{n\cdot\mathrm{kl}\left(\frac{R_n(\theta)}{2U_{\max}},\frac{R(\theta)}{2U_{\max}}\right)\right\}\leq\xi(n). \tag{2.65}$$

Then the result follows from integrating over Θ with respect to π.

First consider any θ such that $R(\theta)=0$ or $R(\theta)=2U_{\max}$. Recall

$$R(\theta)=E\psi(X,Y)1\{a_\theta(X)\neq Y\},$$

$\psi(X,Y)$ can be written $\psi(X,Y)=U(Y,Y,X)-U(-Y,Y,X)\leq 2U_{\max}$, and $\psi(X,Y)>0$ by Assumption 2.2.1. If $R(\theta)=0$ then it follows that we must have $\Pr(a_\theta(X)=Y)=1$ and hence $1\{a_\theta(X_i)\neq Y_i\}=0$ for $i=1,\dots,n$ (a.s.). Hence $R_n(\theta)=0$ in this case (a.s.), so that (2.65) holds. If $R(\theta)=2U_{\max}$, it follows that we must have $\Pr(\psi(X,Y)=2U_{\max})=1$ and $\Pr(a_\theta(X)=Y)=0$, so that now $R_n(\theta)=2U_{\max}$ (a.s.) and again (2.65) holds.

When θ is such that $R(\theta)\notin\{0,2U_{\max}\}$, the proof follows that in Theorem 1 of Maurer (2004) or Lemma 19 in Germain et al. (2015) with minor adjustments. Note that

$$\exp\left\{\lambda D\left(\frac{R(\theta)}{2U_{\max}},\frac{R_n(\theta)}{2U_{\max}}\right)\right\}=\exp\left\{n\cdot\mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^n\frac{\ell(\theta,Y_i,X_i)}{2U_{\max}},\frac{R(\theta)}{2U_{\max}}\right)\right\}$$

is a convex function of $\mathbf{X}=(\ell(\theta,Y_1,X_1)/2U_{\max},\dots,\ell(\theta,Y_n,X_n)/2U_{\max})$ and $\ell(\theta,x,y)/2U_{\max}\in[0,1]$. Then, by Lemma 2.A.2,

$$E\exp\left\{\lambda D\left(\frac{R(\theta)}{2U_{\max}},\frac{R_n(\theta)}{2U_{\max}}\right)\right\}\leq E\exp\left\{n\cdot\mathrm{kl}\left(\frac{1}{n}\sum_{i=1}^n X_i',\frac{R(\theta)}{2U_{\max}}\right)\right\} \tag{2.66}$$

where $X_1',\dots,X_n'$ are iid Bernoulli random variables with success probability $R(\theta)/(2U_{\max})$. Denoting $X'=\sum_{i=1}^n X_i'$,

$$\begin{aligned}
&E\exp\left\{n\cdot\mathrm{kl}\left(\frac{1}{n}X',\frac{R(\theta)}{2U_{\max}}\right)\right\}\\
&\quad=E\left(\frac{\frac{1}{n}X'}{\frac{R(\theta)}{2U_{\max}}}\right)^{X'}\left(\frac{1-\frac{1}{n}X'}{1-\frac{R(\theta)}{2U_{\max}}}\right)^{n-X'}\\
&\quad=\sum_{k=0}^n\Pr\left(X'=k\right)\left(\frac{\frac{k}{n}}{\frac{R(\theta)}{2U_{\max}}}\right)^{k}\left(\frac{1-\frac{k}{n}}{1-\frac{R(\theta)}{2U_{\max}}}\right)^{n-k}\\
&\quad=\sum_{k=0}^n\binom{n}{k}\left(\frac{R(\theta)}{2U_{\max}}\right)^k\left(1-\frac{R(\theta)}{2U_{\max}}\right)^{n-k}\left(\frac{\frac{k}{n}}{\frac{R(\theta)}{2U_{\max}}}\right)^{k}\left(\frac{1-\frac{k}{n}}{1-\frac{R(\theta)}{2U_{\max}}}\right)^{n-k}\\
&\quad=\sum_{k=0}^n\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}=\xi(n)
\end{aligned} \tag{2.67}$$

Therefore (2.65) holds for any $\theta\in\Theta$, completing the proof. ■

Proof of Corollary 2.3.1. We have

$$
\begin{aligned}
&\int_{\Theta}\mathbb{U}_n(\theta)d\hat{\rho}\left(\theta\right)-\int_{\Theta}\mathbb{U}(\theta)d\hat{\rho}\left(\theta\right)\\
&=\frac{1}{n}\sum_{i=1}^{n}\left[U\left(Y_i,Y_i,X_i\right)-EU\left(Y_i,Y_i,X_i\right)\right]+\int_{\Theta}\left[R\left(\theta\right)-R_n\left(\theta\right)\right]d\hat{\rho}\left(\theta\right).
\end{aligned}
$$

Using Hoeffding's inequality, we have

$$
\Pr\left(\frac{1}{n}\sum_{i=1}^{n}\left[U\left(Y_i,Y_i,X_i\right)-EU\left(Y_i,Y_i,X_i\right)\right]>U_{\max}\sqrt{\frac{2\log\frac{2}{\varepsilon}}{n}}\right)\le\frac{\varepsilon}{2}.
$$

Therefore,

$$
\begin{aligned}
&\Pr\left(\int_{\Theta}\mathbb{U}_n(\theta)d\hat{\rho}\left(\theta\right)-\int_{\Theta}\mathbb{U}(\theta)d\hat{\rho}\left(\theta\right)>B_U+B_R\left(\hat{\rho}\right)\right)\\
&\le\Pr\left\{\frac{1}{n}\sum_{i=1}^{n}\left[U\left(Y_i,Y_i,X_i\right)-EU\left(Y_i,Y_i,X_i\right)\right]>B_U\right\}\\
&+\Pr\left\{\int_{\Theta}\left[R\left(\theta\right)-R_n\left(\theta\right)\right]d\hat{\rho}\left(\theta\right)>B_R\left(\hat{\rho}\right)\right\}\\
&\le\frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon,
\end{aligned}
$$

and the result follows. ■

2.A.3 Proofs for Section 2.3.2

Proof of Lemma 2.3.3. By definition and via simple calculations, we have

$$
\begin{aligned}
& D_{\mathrm{KL}}(\rho, \pi) \\
&= -\frac{1}{2} E_{\theta\sim\rho}\left[\log\frac{\det\left(\Sigma_\rho\right)}{\det\left(\Sigma_\pi\right)} + \left(\theta-\mu_\rho\right)'\Sigma_\rho^{-1}\left(\theta-\mu_\rho\right) - \left(\theta-\mu_\pi\right)'\Sigma_\pi^{-1}\left(\theta-\mu_\pi\right)\right] \\
&= -\frac{1}{2}\log\frac{\det\left(\Sigma_\rho\right)}{\det\left(\Sigma_\pi\right)} - \frac{1}{2}\left[M - E_{\theta\sim\rho}\left(\theta-\mu_\rho+\mu_\rho-\mu_\pi\right)'\Sigma_\pi^{-1}\left(\theta-\mu_\rho+\mu_\rho-\mu_\pi\right)\right] \\
&= -\frac{1}{2}\log\frac{\det\left(\Sigma_\rho\right)}{\det\left(\Sigma_\pi\right)} - \frac{1}{2}\left[M - tr\left(\Sigma_\rho\Sigma_\pi^{-1}\right) - \left(\mu_\rho-\mu_\pi\right)'\Sigma_\pi^{-1}\left(\mu_\rho-\mu_\pi\right)\right] \\
&= \frac{1}{2}\left(\mu_\rho-\mu_\pi\right)'\Sigma_\pi^{-1}\left(\mu_\rho-\mu_\pi\right) + \frac{1}{2}\left[tr\left(\Sigma_\rho\Sigma_\pi^{-1}\right) - M\right] - \frac{1}{2}\log\frac{\det\left(\Sigma_\rho\right)}{\det\left(\Sigma_\pi\right)}.
\end{aligned}
$$

■

Proof of Lemma 2.3.4. We have

$$
\begin{aligned}
& \int_\Theta R_n(\theta)\, d\rho(\theta) \\
&= \frac{1}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)E_{\theta\sim\rho}1\left\{Y_i \neq \operatorname{sign}\left[\phi(X_i)'\theta - c(X_i)\right]\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)E_{\theta\sim\rho}1\left\{Y_i\left[\phi(X_i)'\theta - c(X_i)\right] \leq 0\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)E_{\theta\sim\rho}1\left\{\left[Y_i\phi(X_i)'\theta - Y_i c(X_i)\right] \leq 0\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)E_{Z\sim N(0,I_d)}1\left\{\left[Y_i\phi(X_i)'\left(\mu_\rho+\Sigma_\rho^{1/2}Z\right) - Y_i c(X_i)\right] \leq 0\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)\Pr_{Z\sim N(0,I_d)}\left\{Y_i\phi(X_i)'\Sigma_\rho^{1/2}Z \leq Y_i\left[c(X_i) - X_i'\mu_\rho\right]\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\psi(X_i,Y_i)\Phi\left(\frac{Y_i\left[c(X_i) - \phi(X_i)_i'\mu_\rho\right]}{\sqrt{\phi(X_i)'\Sigma_\rho\phi(X_i)}}\right).
\end{aligned}
$$

■

2.A.4 Proofs for Section 2.3.3

Proof of Lemma 2.3.5. The proof is essentially the same as that for Lemma 2.2.2, but we can be more explicit. By definition, we have

$$\begin{aligned}\frac{\rho_{A,\pi}(k)}{\pi(k)} \cdot \frac{d\rho_{A,\pi}\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)} &= \frac{\nu_A(k)}{\sum_{j=1}^{K} \pi(j)\,\nu_A(j)} \cdot \frac{\exp\left(-A\left(k,\theta_{(k)}\right)\right)}{\nu_A(k)} \\ &= \frac{\exp\left(-A\left(k,\theta_{(k)}\right)\right)}{\sum_{j=1}^{K} \pi(j)\,\nu_A(j)}.\end{aligned}$$

Now, using the definition of the KL divergence, we have, for any $\rho \in \mathscr{P}_\pi(\Theta)$:

$$\begin{aligned}&D_{\mathrm{KL}}\left(\rho,\rho_{A,\pi}\right) \\ &= \sum_{k=1}^{K} \left\{ \int_{\Theta_{(k)}} \log\left[\frac{\rho(k)}{\rho_{A,\pi}(k)} \cdot \frac{d\rho\left(\theta_{(k)}|k\right)}{d\rho_{A,\pi}\left(\theta_{(k)}|k\right)}\right] d\rho\left(\theta_{(k)}|k\right)) \right\} \rho(k) \\ &= \sum_{k=1}^{K} \left\{ \int_{\Theta_{(k)}} \log\left\{\frac{\rho(k)}{\pi(k)} \cdot \frac{d\rho\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)} \left[\frac{\rho_{A,\pi}(k)}{\pi(k)} \cdot \frac{d\rho_{A,\pi}\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)}\right]^{-1}\right\} d\rho\left(\theta_{(k)}|k\right)) \right\} \rho(k) \\ &= \sum_{k=1}^{K} \left\{ \int_{\Theta_{(k)}} \left[\log\left(\frac{\rho(k)}{\pi(k)} \cdot \frac{d\rho\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)}\right) - \log\left(\frac{\exp\left(-A\left(k,\theta_{(k)}\right)\right)}{\sum_{j=1}^{K} \pi(j)\,\nu_A(j)}\right)\right] d\rho\left(\theta_{(k)}|k\right)) \right\} \rho(k) \\ &= \sum_{k=1}^{K} \left\{ \int_{\Theta_{(k)}} \log\left[\frac{\rho(k)}{\pi(k)} \cdot \frac{d\rho\left(\theta_{(k)}|k\right)}{d\pi\left(\theta_{(k)}|k\right)}\right] d\rho\left(\theta_{(k)}|k \right\} \rho(k) \\ &+ \sum_{k=1}^{K} \left[\int_{\Theta_{(k)}} A\left(k,\theta_{(k)}\right) d\rho\left(\theta_{(k)}|k\right)\right] \rho(k) + \log\left[\sum_{j=1}^{K} \pi(j)\,\nu_A(j)\right] \\ &= D_{\mathrm{KL}}(\rho,\pi) + \sum_{k=1}^{K} \left[\int_{\Theta_{(k)}} A\left(k,\theta_{(k)}\right) d\rho\left(\theta_{(k)}|k\right)\right] \rho(k) + \log\left[\sum_{j=1}^{K} \pi(j)\,\nu_A(j)\right].\end{aligned}$$

Hence

$$\begin{aligned}&\log\left[\sum_{j=1}^{K} \pi(j)\,\nu_A(j)\right] \\ &= -\left\{D_{\mathrm{KL}}(\rho,\pi) + \sum_{k=1}^{K} \left[\int_{\Theta_{(k)}} A\left(k,\theta_{(k)}\right) d\rho\left(\theta_{(k)}|k\right)\right] \rho(k)\right\} + D_{\mathrm{KL}}\left(\rho,\rho_{A,\pi}\right).\end{aligned}$$

■

Chapter 3

Asymptotic F Test in Regressions With Observations Collected at High Frequency Over Long Span

Abstract

This paper proposes tests of linear hypotheses when the variables may be continuous-time processes with observations collected at a high sampling frequency over a long span. Utilizing series long run variance (LRV) estimation in place of the traditional kernel LRV estimation, we develop easy-to-implement and more accurate F tests in both stationary and nonstationary environments. The nonstationary environment accommodates exogenous regressors that are general semimartingales. Endogeneous regressors are allowed in a nonstationary environment similar to cointegration models in the usual discrete-time setting. The F tests can be implemented in exactly the same way as in the discrete-time setting. The F tests are, therefore, robust to the continuous-time or discrete-time nature of the data. Simulations demonstrate the improved size accuracy and competitive power of the F tests relative to existing continuous-time testing procedures and their improved versions. The F tests are of practical interest as recent work by Chang et al. (2021) demonstrates that traditional inference methods can become invalid and produce spurious results when continuous-time processes are observed on finer grids over a long span.

3.1 Introduction

The advent of high-frequency data poses challenges for classical inference and modeling procedures. For linear regression analysis with observations collected over time, as the grid of observed times becomes finer, continuous-time properties of the underlying processes may conflict with traditional assumptions framed in a discrete-time setting. An immediate concern is the validity of inference procedures when the data generating processes may be continuous-time in nature. Another concern is how we can automate inference procedures so that a researcher can make fewer technical and theoretical modeling decisions. At what sampling frequency should a researcher consider moving to an explicitly continuous-time framework? Should a researcher convert a high-frequency sample into a lower-frequency sample before conducting regression analysis in a discrete-time framework? If continuous-time modeling requires accounting for the sampling frequency, what measurement constitutes a single unit of time? An hour, a day, or a month? Designing trustworthy inference procedures in realistic sample sizes is also a concern.

In this paper, we propose statistical tests that aim to address the above concerns. Recently Chang et al. (2021) considers statistical inference in this setting, highlighting how traditional hypothesis tests can become spurious when observations are collected at a high frequency over a long time span. They show that it is essential to use an autocorrelation-robust variance or long run variance to construct test statistics and make valid inferences. They utilize the continuous-time kernel LRV estimator developed in Lu and Park (2019). Adopting the traditional asymptotic specification that ensures the consistency of the kernel LRV estimator, they show that the test statistics are asymptotically chi-squared. One takeaway from Chang et al. (2021) is that not all kernel-based LRV estimation procedures can be applied without explicitly accounting for the continuous-time environment. A "high-frequency-compatible" bandwidth is desired. Interestingly, the parametric plug-in bandwidth choice of Andrews (1991) is high-frequency-compatible while the nonparametric analogue of Newey and West (1994) is not.

In this paper, we build on Chang et al. (2021) and propose convenient and trustworthy tests in regressions with high-frequency data collected over a long span. We consider both common regressions with stationary regressors and cointegrating regressions with nonstationary regressors. Due to self-normalization, our tests yield valid inferences in the continuous-time setting and would also be valid if the observations were generated from a discrete-time process satisfying standard linear regression assumptions. A practitioner does not have to make any difficult decisions — they can simply use all the observed data, and they can compute the test statistic and perform hypothesis testing in exactly the same way in both the discrete-time and continuous-time settings.

We make several contributions along different dimensions. First, we adopt the more recent fixed-smoothing asymptotic framework. In the discrete-time setting, it is well known that randomness in LRV estimators can lead to significant size distortion of the associated chi-squared tests in finite samples. The same problem is present in the continuous-time setting. By employing the fixed-smoothing asymptotic framework as in Sun (2011, 2013), we show that our test statistics are asymptotically F distributed in both stationary and nonstationary settings. The F approximations capture the randomness of the LRV estimators and are more accurate than the chi-squared approximations.

Second, the asymptotic F theory is based on the series LRV estimator, and in the supplementary appendix, we characterize its asymptotic bias and variance in the high-frequency setting. The series LRV estimator involves projecting the discretized data onto a sequence of orthonormal basis functions and then taking an average of the outer products of the projection coefficients. The number of orthonormal basis functions, denoted by K, is the smoothing parameter in this type of nonparametric variance estimator. Based on the asymptotic bias and variance, we develop a data-driven and automated choice of K in the high-frequency setting. Our rule of selecting K extends that of Phillips (2005), which considers the series LRV estimator in the low-frequency discrete-time setting[1]. Furthermore, we allow for a general class of orthonormal basis functions

[1] Typical examples of low-frequency discrete-time data include monthly and yearly data. The frequency here refers to the sampling frequency, namely the number of times we can draw observations per unit of time. It does not refer to the frequency in the frequency domain that measures the speed that a process completes a cycle.

while Phillips (2005) focuses on sine and cosine functions. See Lazarus et al. (2018) for some practical guidance on using the series LRV estimator with low-frequency discrete-time data.

Third, in a discrete-time cointegrating model, it is common to accommodate endogenous regressors. Following this practice, we allow the regressors to be endogenous in the continuous-time nonstationary setting. This constitutes another departure from Chang et al. (2021) which considers only the case with exogenous regressors. To deal with the endogeneity, we follow Hwang and Sun (2018), but we have to introduce some modifications to facilitate the asymptotic analysis. However, the continuous-time test statistic is computationally identical to the discrete-time statistic in Hwang and Sun (2018), and they are shown to have the same limiting F distribution.

Finally, in the nonstationary setting with exogenous regressors, we establish the asymptotic F distribution for a wider class of regressor processes. The scaled regressor process may converge to a general stochastic process that includes the Brownian motion as a special case. To a great extent, our asymptotic F theory goes beyond its counterpart in the low-frequency discrete-time setting where the nonstationary process is a unit root process and thus converges to a Brownian motion after appropriate normalization.

The class of series LRV estimators is closely related to the class of kernel LRV estimators; see, for example, the discussion in Sun (2011). In essence, a series LRV estimator can be regarded as a kernel LRV estimator with a generalized kernel function. The fixed-K approach adopted here is analogous to the "fixed-b" approach employed in Kiefer and Vogelsang (2005). Fixed-b asymptotics can be developed for the kernel-based test statistics in Chang et al. (2021). However, the limiting distributions are nonstandard and hard to use. They can also be nonpivotal in the nonstationary setting (see Vogelsang and Wagner (2014) for the possible nonpivotality). This provides further justification for the use of series LRV estimation in designing convenient and accurate inference procedures in finite samples.

The rest of the paper is organized as follows. Section 3.2 considers the case where the regressors are stationary, and Section 3.3 considers the nonstationary case with cointegration.

Section 3.4 evaluates the finite sample performances of the proposed F tests, Section 3.5 presents an empirical application, and Section 3.6 concludes. Proofs are given in the appendix. A supplementary appendix develops the MSE-optimal choice of K in the stationary case and recommends a rule of thumb for selecting K. Such a rule is adopted for both the stationary and nonstationary cases in our simulation study.

3.2 The Case with Stationary Regressors

3.2.1 The basic setting

Consider a continuous-time regression of the form

$$Y_t = X_t' \beta_0 + U_t,$$

where each of $Y_t \in \mathbb{R}, X_t \in \mathbb{R}^{d\times 1}$ and $U_t \in \mathbb{R}$ is a continuous-time process for $t \in [0,T]$ with sample paths that are right continuous with left limits (cadlag). We assume that U_t is stationary and $E(U_t|X_s, s \in [0,T]) = 0$ for any $t \in [0,T]$. In this section, we also assume that X_t is a stationary process and defer the case with a nonstationary X_t to Section 3.3. An intercept can be included in X_t in this section.

We do not observe the processes continuously. Instead, for some small sampling interval δ, we observe $\{(x_i, y_i)\}_{i=1}^n$ where

$$x_i = X_{i\delta}, y_i = Y_{i\delta}$$

for $i = 1, \ldots, n$ and $n = T/\delta$. Here, for notational simplicity, we have assumed that T/δ is an integer. The discrete-time sample $\{(x_i, y_i)\}_{i=1}^n$ satisfies

$$y_i = x_i' \beta_0 + u_i, i = 1, 2, \ldots, n,$$

where $u_i = U_{i\delta}$ is unobserved. We are interested in testing $H_0 : R\beta_0 = r$ versus $H_1 : R\beta_0 \neq r$ for some $p \times d$ matrix R with a full row rank p.

Given the discrete sample $\{(x_i, y_i)\}_{i=1}^n$, we estimate β_0 by

$$\hat{\beta}_D = \left(\sum_{i=1}^n x_i x_i'\right)^{-1} \left(\sum_{i=1}^n x_i y_i\right).$$

Our test of H_0 against H_1 is based on the above estimator.

3.2.2 The test statistic

To test whether $R\beta_0$ is equal to r, we often first find the rate of convergence of $\hat{\beta}_D - \beta_0$, establish the asymptotic distribution of a rescaled version of $\hat{\beta}_D - \beta_0$ and then construct the test statistic based on an estimated asymptotic variance. Instead of following these conventional steps, we use heuristic arguments and construct the test statistic directly. The *approximate* variance of $\hat{\beta}_D - \beta_0$ is

$$\left(\sum_{i=1}^n x_i x_i'\right)^{-1} \text{var}\left(\sum_{i=1}^n x_i u_i\right) \left(\sum_{i=1}^n x_i x_i'\right)^{-1}.$$

Based on this *approximate* variance formula, we construct the test statistic

$$F_T = (R\hat{\beta}_D - r)' \left[R\left(\sum_{i=1}^n x_i x_i'\right)^{-1} \widehat{\text{var}}\left(\sum_{i=1}^n x_i \hat{u}_i\right) \left(\sum_{i=1}^n x_i x_i'\right)^{-1} R'\right]^{-1} (R\hat{\beta}_D - r)/p,$$

where $\hat{u}_i = y_i - x_i'\hat{\beta}_D$ and $\widehat{\text{var}}(\sum_{i=1}^n x_i \hat{u}_i)$ is an estimator of the *approximate* variance of $\sum_{i=1}^n x_i u_i$. In the above, dividing by p does not affect the properties of the test.

We use the series estimator for the *approximate* variance. Let $\{\phi_j(\cdot)\}$ be some basis functions on $L^2[0,1]$. The series variance estimator is given by

$$\widehat{\text{var}}\left(\sum_{i=1}^n x_i \hat{u}_i\right) = \frac{1}{K}\sum_{j=1}^K \left[\sum_{i=1}^n \phi_j\left(\frac{i}{n}\right) x_i \hat{u}_i\right]^{\otimes 2}, \tag{3.1}$$

where $a^{\otimes 2} = aa'$ for any vector a and K is a tuning parameter. When the basis functions can be paired naturally, we shall assume that K is even. Note that the basis functions are evaluated at i/n instead of i/T. This is an important point, and our asymptotic theory relies crucially on this construction. We have, therefore, effectively ignored the high-frequency nature of our time series observations that are sampled from continuous time processes. The test statistic is then

$$F_T = (R\hat{\beta}_D - r)' \left\{ R \left(\sum_{i=1}^{n} x_i x_i' \right)^{-1} \frac{1}{K} \sum_{j=1}^{K} \left[\sum_{i=1}^{n} \phi_j \left(\frac{i}{n} \right) x_i \hat{u}_i \right]^{\otimes 2} \left(\sum_{i=1}^{n} x_i x_i' \right)^{-1} R' \right\}^{-1} (R\hat{\beta}_D - r)/p. \tag{3.2}$$

The form of the test statistic is exactly the same as what we would use for a standard regression with discrete time series. Importantly, there is no rescaling by n or T. To construct the test statistic, we can ignore the fact that our observations come from sampling continuous-time processes.

The test statistic F_T takes a self-normalized form. This will become more transparent if we consider the special case that $d = p = 1$ and $K = 1$. In this case, we take $R = 1$ without loss of generality, and the test statistic becomes

$$F_T = \left(\frac{\sum_{i=1}^{n} (x_i u_i)}{\sum_{i=1}^{n} \phi \left(\frac{i}{n} \right) (x_i \hat{u}_i)} \right)^2 := (t_T)^2 .$$

The numerator in the t statistic t_T is a simple sum of $x_i u_i$ while the denominator is a weighted sum of $x_i \hat{u}_i$ with non-diminishing and bounded weights. We expect the numerator and denominator to be of the same order of magnitude no matter what δ is. As a result, t_T and F_T will be stochastically bounded for any sampling interval δ. In this sense, the denominator normalizes the numerator, and thus no additional normalization is needed. This form of self-normalization leads to the invariance of our testing procedure to the sampling interval, which we will develop in greater detail.

3.2.3 Assumptions for the fixed-smoothing asymptotics

We consider the asymptotics along the limiting sequence $\delta \to 0$ and $T \to \infty$. The asymptotics would best reflect the finite sample situation where the observations are collected at a high frequency ($\delta \to 0$) over a long span ($T \to \infty$). To develop the more accurate fixed-smoothing asymptotic approximations, we hold K fixed as $\delta \to 0$ and $T \to \infty$.

The fixed-smoothing asymptotics is developed under several assumptions. First and foremost, for any process $Z = \{Z_t : t \in [0,T]\}$ in this section, we assume that it can be decomposed into a continuous part and a pure-jump part:

$$Z_t = Z_t^c + Z_t^d$$

where $Z_t^d = \sum_{0 \le \tau \le t} \Delta Z_\tau$, $\Delta Z_\tau = Z_\tau - Z_{\tau-}$ and $Z_{\tau-} = \lim_{t \to \tau-} Z_t$. That is, we assume that $\{Z_t\}$ is the sum of a continuous local martingale (i.e., Z_t^c) and a sum of jump terms (i.e., Z_t^d).

Next, we present other technical assumptions and provide some discussion on each.

Assumption 3.2.1 *For $Z_t = X_t U_t$, $X_t' X_t$,*

$$\sum_{0 \le \tau \le T} E \left\| \Delta Z_\tau \right\| = O(T) \text{ as } T \to \infty,$$

where for a matrix M, $\|M\|$ is the Frobenius norm of M.

Assumption 3.2.1 is the same as the first part of Assumption A of Chang et al. (2021). It imposes a restriction on the number and sizes of the jumps in $\{Z_t\}$. The assumption is not stringent and is satisfied, for example, for processes with compound Poisson type jumps if the jump sizes are bounded in L_1 and the jump intensity is proportional to T.

Assumption 3.2.2 *For $j = 1, \dots, K$, each function $\phi_j(\cdot)$ is twice continuously differentiable, and $\int_0^1 \phi_j(t) dt = 0$. Also, $\{\phi_j(\cdot)\}_{j=1}^K$ form an orthonormal set in $L^2[0,1]$.*

Assumption 3.2.2 is very mild and is often maintained in the literature on orthonormal series variance estimation; see, for example, Assumption 1(b) in Sun (2014a). The sine and cosine basis functions (i.e., the Fourier basis functions)

$$\phi_{2j-1}(r) = \sqrt{2}\cos(2\pi jr) \text{ and } \phi_{2j}(r) = \sqrt{2}\sin(2\pi jr) \text{ for } j = 1,\ldots,K/2, \tag{3.3}$$

satisfy this assumption. We will use the Fourier bases in our simulation study. For ease of presentation, we set $\phi_0(\cdot) \equiv 1$, the constant function.

Lemma 3.2.1 *Let Assumptions 3.2.1 and 3.2.2 hold. For $Z_t = X_t U_t$, $X_t' X_t$ and $z_i = Z_{i\delta}$,*

$$\frac{1}{n}\sum_{i=1}^{n} \phi_j\left(\frac{i}{n}\right) z_i = \frac{1}{T}\int_0^T \phi_j\left(\frac{t}{T}\right) Z_t dt + O_p\left(e_{\delta,T}(Z)\right),\ j = 0,1,\ldots,K$$

as $\delta \to 0$ and $T \to \infty$,[2] *where*

$$e_{\delta,T}(Z) = \Delta_{\delta,T}(Z) + \frac{\delta}{T}\sup_{t\in[0,T]} \|Z_t\| + \delta$$

and

$$\Delta_{\delta,T}(Z) = \sup_{\tau,t\in[0,T]}\sup_{|\tau-t|\le\delta} \|Z_\tau^c - Z_t^c\|$$

is the modulus of continuity of the continuous part of Z.

Lemma 3.2.1 shows that the discrete-time average is an approximation to the continuous-time integral with the approximation error controlled by the modulus of continuity of Z, a technical term $\delta \sup_{t\in[0,T]} \|Z_t\|/T$ that captures the edge effects, and the sampling interval δ. In the proof of Lemma 3.2.1, we show that under Assumption 3.2.1, the effect of jumps on the approximation error is of order $O_p(\delta)$.

[2] This should be understood in the following way: $\left\| n^{-1}\sum_{i=1}^{n} \phi_j(i/n) z_i - T^{-1}\int_0^T \phi_j(t/T) Z_t dt \right\| = O_p\left(e_{\delta,T}(Z)\right)$. We use the same convention when O_p or o_p is used in matrix equalities.

Assumption 3.2.3 *For $\{\phi_j\}$ satisfying Assumption 3.2.2,*

$$\frac{1}{T}\int_0^T \phi_j\left(\frac{t}{T}\right) X_t X_t' dt = o_p(1) \textit{ for } j = 1, \ldots, K,$$

and

$$\frac{1}{T}\int_0^T X_t X_t' dt = S + o_p(1)$$

for a positive definite matrix S as $T \to \infty$.

To understand the assumption, let X_{tk} be the k-th element of X_t. Suppose X_t is stationary and $E|X_{tk}X_{tl}X_{sk}X_{sl}| < \infty$ for any $k,l = 1,2,\ldots,d$ and any $t,s \in [0,T]$. Assume further that $cov(X_{tk}X_{tl}, X_{sk}X_{sl}) = f_{kl}(t-s)$ for some bounded function $f_{kl}(\cdot)$ satisfying $f_{kl}(\tau) \to 0$ as $|\tau| \to \infty$. Then, by the Fubini–Tonelli theorem,

$$E\frac{1}{T}\int_0^T \phi_j\left(\frac{t}{T}\right) X_t X_t' dt = E\left(X_t X_t'\right) \cdot \frac{1}{T}\int_0^T \phi_j\left(\frac{t}{T}\right) dt = E\left(X_t X_t'\right)\int_0^1 \phi_j(r)\, dr,$$

for all $j = 0,1,\ldots,K$. By the Fubini–Tonelli theorem and the dominated convergence theorem,

$$\begin{aligned}
& var\left(\frac{1}{T}\int_0^T \phi_j\left(\frac{t}{T}\right) X_{ti}X_{tk} dt\right) \\
&= \frac{1}{T^2}\int_0^T\int_0^T \phi_j\left(\frac{t}{T}\right)\phi_j\left(\frac{s}{T}\right) cov(X_{ti}X_{tk}, X_{si}X_{sk}) dtds \\
&= \frac{1}{T^2}\int_0^T\int_0^T \phi_j\left(\frac{t}{T}\right)\phi_j\left(\frac{s}{T}\right) f_{kl}(t-s)\, dtds \\
&= \int_0^1\int_0^1 \phi_j(s)\,\phi_j(t)\, f_{kl}(T(t-s))\, dtds \to 0
\end{aligned}$$

for $j = 0,1,\ldots,K$. Hence Assumption 3.2.3 holds for $S = E\left(X_t X_t'\right)$.

Assumption 3.2.4 *For $\{\phi_j\}$ satisfying Assumption 3.2.2,*

$$\frac{1}{\sqrt{T}}\int_0^T \phi_j\left(\frac{t}{T}\right) X_t U_t dt \Rightarrow \Omega^{1/2}\int_0^1 \phi_j(r)\, dW_d(r) \textit{ jointly for } j = 0,1,2,\ldots,K$$

as $\delta \to 0$ and $T \to \infty$, where $W_d(r)$ is the $d \times 1$ standard Brownian motion process,

$$\Omega = \lim_{T\to\infty} \operatorname{var}\left(\frac{1}{\sqrt{T}}\int_0^T X_t U_t dt\right) = \int_{-\infty}^{\infty} \Gamma_{XU}(\tau)\, d\tau,$$

$\Gamma_{XU}(\tau) = E\left[X_t U_t U_{t-\tau} X'_{t-\tau}\right]$*, and $\Omega^{1/2}$ is a matrix square root of Ω so that $\Omega^{1/2}\left(\Omega^{1/2}\right)' = \Omega$.*

Assumption 3.2.4 is a multivariate CLT in the continuous-time setting. As in the discrete time setting, there is a large body of literature on CLT's for additive functionals in a continuous time setting. For example, Rozanov (1960) establishes a CLT for additive functionals such as $T^{-1/2}\int_0^T \phi_j(t/T) X_t U_t dt$. The sufficient conditions, which include a mixing condition and a moment condition, are similar to those in the discrete time setting.

If a functional CLT (FCLT) holds such that $T^{-1/2}\int_0^{[Tr]} X_t U_t dt \Rightarrow \Omega^{1/2} W_d(r)$, then using integration by parts and the continuous mapping theorem, we can show that Assumption 3.2.4 holds. Sufficient conditions for the FCLT for the class of functions of continuous-time stationary ergodic Markov processes can be founded in Bhattacharya (1982). For more discussions, see Equations 1–3 and remarks in Section 2 of Lu and Park (2019). Note that an FCLT is stronger than necessary, but the gap between an FCLT and the above multivariate CLT may be of theoretical interest only. Here we only need a multivariate CLT. This is an advantage of using a series LRV estimator. If we use a kernel LRV estimator, then an FCLT is needed for developing fixed-smoothing asymptotics.

Assumption 3.2.5 *(i) $\sqrt{T} e_{\delta,T}(XU) = o_p(1)$ and (ii) $e_{\delta,T}(XX') = o_p(1)$.*

Assumption 3.2.5 is the same as Assumption D1 of Chang et al. (2021). Assumption 3.2.5(i) holds if $\sqrt{T}\delta = o(1)$, $\sqrt{T}\Delta_{\delta,T}(XU) = o_p(1)$ and $\sup_{t\in[0,T]} \|XU\| = o_p(\sqrt{T}/\delta)$. The first condition, namely $\sqrt{T}\delta = o(1)$, requires that $\delta \to 0$ fast enough as $T \to \infty$, that is, the continuous-time process has to be sampled frequently enough. The second condition, namely $\sqrt{T}\Delta_{\delta,T}(XU) = o_p(1)$, requires that the continuous part of $\{X_t U_t\}$ does not fluctuate too much over the sampling intervals of length δ. Using the moment bounds in Fischer and Nappo (2009) and the Markov inequality, we can obtain that

$$\Delta_{\delta,T}\left(XU\right)=O_p\left[\left(\delta\log\frac{2T}{\delta}\right)^{1/2}\right]$$

if $\left(X_tU_t\right)^c$ is an Ito process whose drift and diffusion coefficients satisfy some mild conditions. So $\sqrt{T}\Delta_{\delta,T}\left(XU\right)=o_p\left(1\right)$ if $T\delta\log\left(T/\delta\right)=o\left(1\right)$. The third condition, namely $\sup_{t\in[0,T]}\left\|XU\right\|=o_p(\sqrt{T}/\delta)$, requires that the maximum value of the process $\{X_tU_t\}$ over $[0,T]$ does not explode too quickly as T grows. For example, if $\sup_{t\in[0,T]}\left\|XU\right\|=O_p\left(T\right)$ and $\sqrt{T}\delta=o\left(1\right)$, then $\sup_{t\in[0,T]}\left\|XU\right\|=O_p\left(T\right)=O_p(\sqrt{T}\delta\cdot\sqrt{T}/\delta)=o_p(\sqrt{T}/\delta)$ and the third condition holds. Assumption 3.2.5(ii) is of the same form as Assumption 3.2.5(i). With some obvious modifications, our discussions on Assumption 3.2.5(i) can be applied to Assumption 3.2.5(ii).

3.2.4 Fixed-smoothing asymptotics

Define

$$\hat{\beta}_C=\left[\int_0^T X_tX_t'dt\right]^{-1}\left[\int_0^T X_tY_tdt\right],$$

which is the least-square analogue of $\hat{\beta}_D$ in the space $L^2\left[0,T\right]$ using the continuous-time data $\left\{\left(X_t,Y_t\right),t\in\left[0,T\right]\right\}$. $\hat{\beta}_C$ is not feasible, and we use it only as a benchmark for comparison.

We first show that $\sqrt{T}[\hat{\beta}_D-\beta]$ and $\sqrt{T}[\hat{\beta}_C-\beta]$ are asymptotically equivalent. Letting $Z_t=X_tU_t$ and $j=0$ in Lemma 3.2.1, we have

$$\frac{1}{n}\sum_{i=1}^{n}x_iu_i=\frac{1}{T}\int_0^T X_tU_tdt+O_p\left(e_{\delta,T}\left(XU\right)\right).$$

Multiplying the above equation by $\sqrt{T}$, we obtain

$$\frac{1}{\Lambda\left(n,\delta\right)}\sum_{i=1}^{n}x_iu_i=\frac{1}{\sqrt{T}}\int_0^T X_tU_tdt+o_p\left(1\right),$$

where $\Lambda(n,\delta) = \sqrt{n/\delta}$ and we have used Assumption 3.2.5(i).

Using Lemma 3.2.1 with $Z_t = X_t X_t'$ and $j = 0$ and Assumption 3.2.5(ii), we have

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' = \frac{1}{T}\int_0^T X_t X_t' dt + o_p(1).$$

Hence,

$$\begin{aligned}
\sqrt{T}\left[\hat{\beta}_D - \beta_0\right] &= \left[n/\Lambda(n,\delta)\right]\left[\hat{\beta}_D - \beta_0\right] \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n} x_i u_i\right) \\
&= \left(\frac{1}{T}\int_0^T X_t X_t' dt\right)^{-1}\frac{1}{\sqrt{T}}\int_0^T X_t U_t dt + o_p(1) \\
&= \sqrt{T}(\hat{\beta}_C - \beta_0) + o_p(1).
\end{aligned}$$

The above derivations show that Assumptions 3.2.1 and 3.2.5 ensure that $\sqrt{T}(\hat{\beta}_D - \beta_0)$ and $\sqrt{T}(\hat{\beta}_C - \beta_0)$ are asymptotically equivalent. Invoking Assumptions 3.2.3 and 3.2.4, we obtain the asymptotic distribution of $\sqrt{T}(\hat{\beta}_D - \beta)$. We present this and another key result, which requires Assumption 3.2.2, in the lemma below.

Lemma 3.2.2 *Let Assumptions 3.2.1–3.2.5 hold. Then*

$$\sqrt{T}(\hat{\beta}_D - \beta_0) = \sqrt{T}(\hat{\beta}_C - \beta_0) + o_p(1) \Rightarrow S^{-1}\Omega^{1/2}W_d(1)$$

and

$$\frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)x_i\hat{u}_i \Rightarrow \Omega^{1/2}\int_0^1 \phi_j(r)\,dW_d(r)$$

jointly for $j = 1, 2, \ldots, K$.

Lemma 3.2.2 shows that $\hat{\beta}_D$ converges to β_0 at the rate of $\sqrt{T}$. For high-frequency data sampled from a continuous-time process, the effective sample size is the time span T rather than the number of observations n. We do not obtain the rate of $\sqrt{n}$, which is the typical rate for the discrete-time data with a fixed sampling interval (e.g., δ is fixed to be 1) and n weakly dependent observations. The difference can be traced back to the unusual rate in the weak convergence result:

$$\frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n} x_i u_i \Rightarrow \Omega^{1/2} W_d(1).$$

Because $\{x_i u_i\}$ becomes highly correlated as $\delta \to 0$, in order to obtain a well-defined weak limit, we need to normalize the sum $\sum_{i=1}^{n} x_i u_i$ by $\Lambda(n,\delta) := \sqrt{n/\delta}$, which is larger than the usual normalization factor $\sqrt{n}$ by an order of magnitude.

Using Lemma 3.2.2, we have, under the null hypothesis:

$$\begin{aligned}
F_T &= \delta\Lambda(n,\delta)(R\hat{\beta}_D - r)' \\
&\times \left[R\left(\frac{1}{\delta\Lambda(n,\delta)^2}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\frac{1}{K}\sum_{j=1}^{K}\left[\frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)x_i\hat{u}_i\right]^{\otimes 2}\left(\frac{1}{\delta\Lambda(n,\delta)^2}\sum_{i=1}^{n} x_i x_i'\right)^{-1} R'\right]^{-1} \\
&\times \delta\Lambda(n,\delta)(R\hat{\beta}_D - r)/p \\
&\Rightarrow [RS^{-1}\Omega^{1/2}W_d(1)]'\left\{RS^{-1}\Omega^{1/2}\frac{1}{K}\sum_{j=1}^{K}\left[\int_0^1 \phi_j(r)\,dW_d(r)\right]^{\otimes 2}\Omega^{1/2}S^{-1}R'\right\}^{-1} RS^{-1}\Omega^{1/2}W_d(1)/p.
\end{aligned}$$

In the above, rescalings by $\delta\Lambda(n,\delta)$, $1/\Lambda(n,\delta)$ or $1/(\delta\Lambda(n,\delta)^2)$ in the first equality are for theoretical arguments only. In practice, the test statistic F_T is computed according to the definition in (3.2) without using any rescaling.

Note that $RS^{-1}\Omega^{1/2}W_d(r) \overset{d}{=} \left[RS^{-1}\Omega S^{-1}R'\right]^{1/2} W_p(r)$ for a $p \times 1$ standard Brownian motion process $W_p(\cdot)$ and that $RS^{-1}\Omega S^{-1}R'$ is of a full rank. We have

$$F_T \Rightarrow [W_p(1)]'\left\{\frac{1}{K}\sum_{j=1}^{K}\left[\int_0^1 \phi_j(r)\,dW_p(r)\right]^{\otimes 2}\right\}^{-1} W_p(1)/p.$$

Under Assumption 3.2.2, $\left[\int_0^1 \phi_j(r)\,dW_p(r)\right]^{\otimes 2}$ is iid Wishart distributed. The above limiting distribution is equal to Hotelling's T^2 distribution. In view of the relationship between the T^2 and F distributions (e.g., Bilodeau and Brenner (2010)), we have the following theorem.

Theorem 3.2.1 *Let Assumptions 3.2.1 – 3.2.5 hold. Then, for a fixed $K \geq p$,*

$$F_T \Rightarrow \frac{K}{K-p+1} F_{p,K-p+1},$$

where $F_{p,K-p+1}$ is the F distribution with degrees of freedom p and $K-p+1$.

If we use the OLS variance estimator that ignores the autocorrelation, we would construct the test statistic as follows

$$F_{T,OLS} = \left(R\hat{\beta}_D - r\right)' \times \left[R\hat{\sigma}_u^2 \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} R'\right]^{-1} \left(R\hat{\beta}_D - r\right)/p,$$

where $\hat{\sigma}_u^2 = n^{-1}\sum_{i=1}^{n} \hat{u}_i^2$ is an estimator of the variance σ_u^2 of U_t. Then

$$\begin{aligned} \delta F_{T,OLS} &= \sqrt{T}\left(R\hat{\beta}_D - r\right)' \times \left[R\hat{\sigma}_u^2 \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} R'\right]^{-1} \sqrt{T}\left(R\hat{\beta}_D - r\right)/p \\ &\Rightarrow \left[RS^{-1}\Omega^{1/2}W_d(1)\right]' \times \left[\sigma_u^2 RS^{-1}R'\right]^{-1} \left[RS^{-1}\Omega^{1/2}W_d(1)\right]/p. \end{aligned}$$

So, as $\delta \to 0$, $F_{T,OLS} \to \infty$ with probability approaching one. Consequently, using $F_{T,OLS}$ for inference can lead to the spurious finding of a significant relationship that does not actually exist. See Chang et al. (2021) for more details. Such a result is also related to the following result in Sun (2004): the t-statistic can be made convergent in a spurious regression when high-order autocorrelations are properly accounted for.

To illustrate the key difference between the variance estimators underlying F_T and $F_{T,OLS}$, consider the special case with $K = d = p = 1$. Then the ratio of the autocorrelation robust variance estimator to the OLS variance estimator is

$$\begin{aligned}\frac{\left[\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)(x_i\hat{u}_i)\right]^2}{\hat{\sigma}_u^2\sum_{i=1}^{n}x_i^2} &= \frac{\Lambda(n,\delta)^2}{n}\frac{\left[\frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)x_i\hat{u}_i\right]^2}{\hat{\sigma}_u^2\frac{1}{n}\sum_{i=1}^{n}x_i^2}\\ &= \frac{1}{\delta}\cdot\frac{\left[\frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)x_i\hat{u}_i\right]^2}{\hat{\sigma}_u^2\frac{1}{n}\sum_{i=1}^{n}x_i^2}.\end{aligned}$$

Note that the second factor converges to a nondegenerate distribution. So the ratio will diverge at the rate of $1/\delta$. That is, by ignoring the high-order autocorrelations of x_iu_i, especially when δ is small, the OLS variance estimator under-estimates the true variation of the OLS estimator by a factor of $1/\delta$. This explains why F_T is stochastically bounded while $F_{T,OLS}$ explodes as $\delta\to 0$ and $T\to\infty$.

To implement the F test, we need to choose K. Ideally, we want to select K to tradeoff the type I and type II errors of the F test, but this is well beyond the scope of this paper. In the supplementary appendix, we consider the infeasible LRV estimator

$$\hat{\Omega}^* = \frac{1}{K}\sum_{j=1}^{K}\left[\frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)(x_iu_i)\right]^{\otimes 2}, \tag{3.4}$$

and establish its asymptotic bias and variance under both a fixed K and a growing K (i.e., $K\to\infty$). Note that $\hat{\Omega}^*$ can be regarded as an infeasible version of the variance estimator in (3.1) (after a normalization), as $\{u_i\}$ are not observed. Based on the asymptotic mean square error (MSE) of $\hat{\Omega}^*$, we obtain the MSE-optimal choice of K given in (3.28) in the supplementary appendix.

We recommend using a parametric AR(1) plug-in approach to obtain a data-driven value for K. More specifically, we fit an AR(1) model to each component $z_{i,j}$ of $\{z_i := x_iu_i\}_{i=1}^{n}$:

$$z_{i,j} = \rho_j z_{i-1,j} + e_{zj} \text{ for } j = 1,2,\ldots,d$$

with the AR parameter and error variance estimated by

$$\hat{\rho}_j = \frac{\sum_{i=2}^{n} z_{i,j} z_{i-1,j}}{\sum_{i=2}^{n} z_{i-1,j}^2} \text{ and } \hat{\sigma}_j^2 = \frac{1}{n} \sum_{i=2}^{n} \left(z_{i,j} - \hat{\rho}_j z_{i-1,j} \right)^2 .$$

On the basis of the above plug-in estimates, we compute

$$\hat{\kappa}_D = \frac{1}{8c_{\phi,2}^2} \left(\sum_{j=1}^{d} \frac{\hat{\rho}_j^2 \hat{\sigma}_j^4}{\left(1-\hat{\rho}_j\right)^8} \right)^{-1} \left(\sum_{j=1}^{d} \frac{\hat{\sigma}_j^4}{\left(1-\hat{\rho}_j\right)^4} \right)$$

and then let

$$\hat{K}_D = \hat{\kappa}_D^{1/5} n^{4/5}. \tag{3.5}$$

Note that $c_{\phi,2} = \pi^2/6$ when the Fourier basis functions in (3.3) are used. Our approach is similar to what is proposed in Andrews (1991), but there is an important difference. We do not follow Andrews (1991) and truncate the estimator of the AR coefficient from below, as otherwise we will not have a "high-frequency compatible" choice of K. See the supplementary appendix for more discussions and details.

To conclude this section, we have shown that, in the stationary case, we do not need to change our estimation and inference methods to account for the fact that our observations are collected at a high frequency with the sampling interval δ going to zero. We can use exactly the same approach as we would do in the case with discrete-time observations where the time distance between neighboring observations is fixed: the test statistic is constructed in the same way, and the smoothing parameter is chosen in the same way. We do not need to choose a unit of time to measure the sampling duration. The only caveat is that we should use a parametric AR(1) plug-in to obtain the data-driven smoothing parameter. Using the nonparametric approach of Newey and West (1994) will lead to a sub-optimal rate for the smoothing parameter. See Chang et al. (2021) for the details.

3.3 The Nonstationary Case

3.3.1 Exogenous Regressors

In this subsection, we consider linear hypothesis testing for cointegrating regressions in the continuous-time setting. The model is

$$Y_t = \alpha_0 + X_t'\beta_0 + U_{0t} \tag{3.6}$$

where $X_t \in \mathbb{R}^{d\times 1}$ is a nonstationary process, $U_{0t} \in \mathbb{R}$ is a stationary process, $\{X_t\}$ and $\{U_{0t}\}$ are independent.[3] As in the case with stationary regressors, only a discrete set of points $\{x_i = X_{i\delta}, y_i = Y_{i\delta}\}_{i=1}^n$ are observed. The discrete-time model is

$$y_i = \alpha_0 + x_i'\beta_0 + u_{0i}$$

where $u_{0i} = U_{0,i\delta}$. The object of interest is the slope parameter β_0, and we aim at testing $H_0 : R\beta_0 = r$ against $H_1 : R\beta_0 \neq r$ where $R \in \mathbb{R}^{p\times d}$ is of rank p. Note that here we single the intercept out of the slope parameter, and the hypothesis of interest involves only the slope parameter.

We consider the same limiting experiment where $\delta \to 0$ and $T \to \infty$ for a fixed K.

Assumption 3.3.1 *For $e_{\delta,T}(U_{0\cdot})$ defined in the same way as in Lemma 3.2.1,*

$$\sum_{0\leq\tau\leq T} E\left|\Delta U_{0\tau}\right| = O(T) \text{ and } e_{\delta,T}(U_{0\cdot}) = o_p(1).$$

The above assumption is similar to Assumptions 3.2.1 and 3.2.5(i). It ensures that

$$\Lambda(n,\delta)^{-1}\sum_{i=1}^n u_{0i} = T^{-1/2}\int_0^T U_{0t}dt + o_p(1).$$

[3] We use U_{0t} instead of U_t to denote the error process because in the next subsection we will use U_t to denote $(U_{0t}', U_{xt}')'$. We shall use $U_{0\cdot}$ to denote $\{U_{0t} : t \in [0,T]\}$.

Assumption 3.3.2 *For a sequence of $d \times d$ diagonal matrices (Λ_T) with diverging diagonal elements*

$$\begin{pmatrix} \Lambda_T^{-1} X_{Tr} \\ T^{-1/2} \int_0^{Tr} U_{0s} ds \end{pmatrix} \Rightarrow \begin{pmatrix} X^{\circ}(r) \\ \sigma_0 W_0(r) \end{pmatrix} \text{ for } \sigma_0 > 0 \text{ and } r \in (0,1]$$

as $T \to \infty$, where $X^{\circ}(\cdot)$ is a continuous (a.s.) semimartingale, $W_0(\cdot)$ is standard Brownian motion, and $X^{\circ}(\cdot)$ and $W_0(\cdot)$ are independent.

The weak convergence in Assumption 3.3.2 is defined on $\mathbb{D}^{d+1}[0,1]$, the space of cadlag functions from $[0,1]$ to $\mathbb{R}^{(d+1)\times 1}$ endowed with the Skorokhod topology. The assumption is the continuous-time analogue of the traditional invariance principles. It is similar to Assumption C2 in Chang et al. (2021) which points out that the assumption is satisfied for a wide class of continuous-time processes. For general null recurrent diffusions and jump diffusions, Kim and Park (2017) provides sufficient conditions under which $\Lambda_T^{-1} X_{Tr} \Rightarrow X^{\circ}(r)$. As discussed after Assumption 3.2.4, Lu and Park (2019) provides sufficient conditions under which $T^{-1/2} \int_0^{Tr} U_{0s} ds \Rightarrow \sigma_0 W_0(r)$.

For $j = 1, \ldots, K$, let

$$\eta_j = \int_0^1 \phi_j(r) X^{\circ}(r) dr,$$

and

$$\eta = (\eta_1, \ldots, \eta_K)' \in \mathbb{R}^{K \times d}.$$

Assumption 3.3.3 *With probability one, $\eta'\eta$ is of full rank d.*

Assumption 3.3.3 requires that, with probability one, the $L^2[0,1]$ projection coefficients of components of X° in the directions $\phi_j, j = 1, \ldots, K$, form d linearly independent vectors. For a given choice of $\{\phi_j\}_{j=1}^K$, such as the first K Fourier basis functions given in (3.3), this is satisfied by virtually all continuous-time processes used in practice when K is large enough.

Now we detail the testing procedure. Assume that $K \geq d+1$. The testing steps are as follows:

1. Create the transformed data $\{\mathbb{W}_j^y, \mathbb{W}_j^x\}_{j=1}^K$ where

$$\mathbb{W}_j^y = \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_j\left(\frac{i}{n}\right) y_i,\ \mathbb{W}_j^x = \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_j\left(\frac{i}{n}\right) x_i. \tag{3.7}$$

   Denote the matrix forms of transformed data by

$$\underset{K\times 1}{\mathbb{W}^y = (\mathbb{W}_1^y, \ldots, \mathbb{W}_K^y)'},\ \underset{K\times d}{\mathbb{W}^x = (\mathbb{W}_1^x, \ldots, \mathbb{W}_K^x)'}.$$

2. Regress $\mathbb{W}^y$ on $\mathbb{W}^x$ without an intercept by OLS. This yields the transformed OLS estimator $\hat{\beta}_{TOLS}$ and the residual vector $\hat{\mathbb{W}}^{u_0}$:

$$\hat{\beta}_{TOLS} = \left(\mathbb{W}^{x\prime}\mathbb{W}^x\right)^{-1} \mathbb{W}^{x\prime}\mathbb{W}^y,\ \hat{\mathbb{W}}^{u_0} = \mathbb{W}^y - \mathbb{W}^x\hat{\beta}_{TOLS}. \tag{3.8}$$

3. To test $H_0 : R\beta_0 = r$, we calculate the following test statistic

$$F_{TOLS} = \frac{1}{\hat{\sigma}_0^2}(R\hat{\beta}_{TOLS} - r)' \left[R\left(\mathbb{W}^{x\prime}\mathbb{W}^x\right)^{-1} R'\right]^{-1} (R\hat{\beta}_{TOLS} - r)/p, \tag{3.9}$$

   where

$$\hat{\sigma}_0^2 = \frac{1}{K}\sum_{j=1}^K (\hat{\mathbb{W}}_j^{u_0})^2 = \frac{1}{K}\hat{\mathbb{W}}^{u_0\prime}\hat{\mathbb{W}}^{u_0}. \tag{3.10}$$

Define

$$\mathbb{W}_j^{u_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi_j\left(\frac{i}{n}\right) u_{0i},\ \underset{K\times 1}{\mathbb{W}^{u_0} = (\mathbb{W}_1^{u_0}, \ldots, \mathbb{W}_K^{u_0})'}.$$

For $j = 1, \ldots, K$, let

$$\nu_j = \sigma_0 \int_0^1 \phi_j(r) dW_0(r),$$

and

$$\nu = (\nu_1, \ldots, \nu_K)' \in \mathbb{R}^{K\times 1}.$$

The following lemma establishes the weak limits of $\mathbb{W}^x$, $\mathbb{W}^{u_0}$, and $\hat{\beta}_{TOLS}$.

Lemma 3.3.1 *Let Assumptions 3.2.2, 3.3.1–3.3.3 hold. Then, as $\delta \to 0$ and $T \to \infty$,*

(*a*) $(n^{-1/2}\mathbb{W}^{x}\Lambda_T^{-1}, \sqrt{\delta}\mathbb{W}^{u_0}) \Rightarrow (\eta, \nu)$;

(*b*) $\sqrt{T}\Lambda_T(\hat{\beta}_{TOLS} - \beta_0) \Rightarrow (\eta'\eta)^{-1}(\eta'\nu)$.

Let $R(\ell,\cdot)$ and r_ℓ be the ℓ-th rows of R and r, respectively. Since we do not require that all elements of (X_{Tr}) converge at the same rate, the rate of convergence of $R(\ell,\cdot)\hat{\beta}_{TOLS}$ depends on the element of $\hat{\beta}_{TOLS}$ that has the slowest rate of convergence among those elements appearing in the ℓ-th restriction. To capture this, for $\ell = 1,\ldots,p$, we define the sets

$$\mathscr{I}_\ell := \{j : \text{ for } j \in \{1,2,\ldots,d\} \text{ such that } R(\ell,j) \neq 0\},$$

which consists of the indices of the coefficients that appear in the ℓ-th restriction. When T is large enough, the rate of convergence of $R(\ell,\cdot)\hat{\beta}_{TOLS}$ is given by $\sqrt{T}\min_{j\in\mathscr{I}_\ell}\Lambda_T(j,j)$. Let

$$\tilde{\Lambda}_T = \operatorname{diag}\left(\min_{j\in\mathscr{I}_1}\Lambda_T(j,j),\ldots,\min_{j\in\mathscr{I}_p}\Lambda_T(j,j)\right),$$

which is a $p \times p$ diagonal matrix.[4] Then $\lim_{T\to\infty}\tilde{\Lambda}_T R \Lambda_T^{-1} = R_\circ$ for a matrix $R_\circ \in \mathbb{R}^{p\times d}$ whose (ℓ,j)-th element $R_\circ(\ell,j)$ is equal to

$$R_\circ(\ell,j) = \lim_{T\to\infty}\tilde{\Lambda}_T(\ell,\ell)R(\ell,j)/\Lambda_T(j,j) = R(\ell,j)\lim_{T\to\infty}\left[\min_{m\in\mathscr{I}_\ell}\Lambda_T(m,m)/\Lambda_T(j,j)\right]. \quad (3.11)$$

That is, $R_\circ$ is the same as R after we zero out the elements in each row of R for which the corresponding coefficients can be estimated at a faster rate than the slowest rate for the coefficients involved in this row. We require that $R_\circ$ be of row rank p, a condition that is clearly satisfied when there is no heterogeneity in the rates of convergence, for example, $R_\circ = R$ when Λ_T is a scalar matrix.

[4] $\min_{j\in\mathscr{I}_\ell}\Lambda_T(j,j)$ should be interpreted as the minimum of $\Lambda_T(j,j)$ over $j \in \mathscr{I}_\ell$ when T is large enough.

Theorem 3.3.1 *Let Assumptions 3.2.2, 3.3.1–3.3.3 hold. If $K \geq d+1$ and $\lim_{T\to\infty} \tilde{\Lambda}_T R \Lambda_T^{-1}$ is of rank p, then*

$$F_{TOLS} \Rightarrow \frac{K}{K-d} \cdot F_{p,K-d},$$

where $F_{p,K-d}$ is the F distribution with degrees of freedom p and $K-d$.

Note that the asymptotic F theory does not depend on the specific form of the limiting process $X^{\circ}(\cdot)$. In the proof of the theorem, we show that the asymptotic distribution conditional on $X^{\circ}(\cdot)$ is an F distribution, which does not depend on the conditioning process $X^{\circ}(\cdot)$. Hence, the asymptotic distribution is also the F distribution unconditionally. Asymptotic F theory in a regression with nonstationary and exogenous regressors has been recently developed in Sun (2022) for discrete time series. Since the limiting process $X^{\circ}(\cdot)$ can be highly nonstandard and goes beyond what has been considered in Sun (2022), Theorem 3.3.1 has widened the applicability of the asymptotic F theory. See Kim and Park (2017) for the nonstandard forms that $X^{\circ}(\cdot)$ can take when $\{X_t\}$ is a null recurrent diffusion process.

To implement the F test, we need to choose K. Note that the variance estimator in (3.10) takes a form similar to that in the stationary case. The infeasible variance estimator can be written as

$$\hat{\sigma}_0^2 = \frac{1}{K}\sum_{j=1}^{K}\left[\frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)u_{0i}\right]^2,$$

which can be compared with $\hat{\Omega}^*$ defined in (3.1).

As a practical rule of thumb, we can adapt the data-driven procedure in the stationary case and proceed as follows:

1. Estimate the model $y_i = \alpha_0 + x_i'\beta_0 + u_{0i}$ by OLS to obtain the residual

$$\hat{u}_{0i} = y_i - \hat{\alpha}_{OLS} - x_i'\hat{\beta}_{OLS}.$$

2. On the basis of $\{\hat{u}_{0i}\}$, use the series method to estimate the long run variance of $\{u_{0i}\}$,

computing the AR(1) data-driven $\hat{K}_D$ using the formula in (3.5).

3. Let $\hat{K}^* = \max(\hat{K}_D, d+3)$ and use $\hat{K}^*$ to construct the transformed regression. Taking the maximum between $\hat{K}_D$ and $d+3$ ensures that the limiting F distribution has a finite mean.

4. Compute the F test statistic in the TOLS regression. Perform the asymptotic F test using $\frac{\hat{K}^*}{\hat{K}^*-d} \cdot F_{p,\hat{K}^*-d}$ as the reference distribution.

We note in passing that an asymptotic F theory may also be developed based on the usual OLS estimator in step 1 above rather than the transformed OLS estimator, but then a series variance estimator with judiciously crafted basis functions has to be used. See Sun (2022) for more details in the discrete-time setting. We will not pursue this extension and choose to use a transformed regression, which can be regarded as a special case of the transformed and augmented regression in the next subsection. Hwang and Sun (2018) provides some discussion on the advantages of the transformed approach, including its robustness to contaminations whose energy is concentrated at high frequencies in the frequency domain.

3.3.2 Endogenous Regressors

We consider the same model $Y_t = \alpha_0 + X_t'\beta_0 + U_{0t}$ as in the previous subsection, but we now allow $\{X_t\}$ to be endogenous. The cost of admitting endogeneity comes in the form of less flexibility for the data generating process of the weak limit of $\Lambda_T^{-1}X_{Tr}$, $r \in [0,1]$. Namely, we require that $\Lambda_T^{-1} = T^{-1/2}I_d$ and that the limiting process be Brownian motion. As we discuss shortly, this requirement is a natural adaptation of the discrete time literature on inference in cointegrating regressions. For example, it is similar to the discrete time framework adopted in Vogelsang and Wagner (2014) and Hwang and Sun (2018). It is an open question whether an asymptotic F theory can still be developed for other forms of nonstationarity. As before, we only observe a discrete set of points $\{(x_i, y_i)\}_{i=1}^n$ satisfying $y_i = \alpha_0 + x_i'\beta_0 + u_{0i}$. Again we want to test $H_0 : R\beta_0 = r$ against $H_1 : R\beta_0 \neq r$.

We maintain Assumption 3.3.1 regarding the stationary process $\{U_{0t}\}$ but now allow for some forms of dependence between $\{X_t\}$ and $\{U_{0t}\}$. Towards this end, the assumption below is similar to and replaces Assumption 3.3.2.

Assumption 3.3.4 *As $T \to \infty$, the following functional central limit theorem holds:*

$$\begin{pmatrix} \frac{1}{\sqrt{T}} \int_0^{Tr} U_{0s} ds \\ \frac{1}{\sqrt{T}} X_{Tr} \end{pmatrix} \Rightarrow \begin{pmatrix} B_0(r) \\ B_x(r) \end{pmatrix} := \Omega^{1/2} \begin{pmatrix} W_0(r) \\ W_x(r) \end{pmatrix} \text{ for } r \in [0,1]$$

where $\Omega^{1/2}\left(\Omega^{1/2}\right)' = \Omega$,

$$\Omega = \begin{pmatrix} \underset{1\times 1}{\sigma_0^2} & \underset{1\times d}{\sigma_{0x}} \\ \underset{d\times 1}{\sigma_{x0}} & \underset{d\times d}{\Omega_{xx}} \end{pmatrix},$$

and $W_0(\cdot)$ and $W_x(\cdot)$ are independent standard Brownian motions.

The weak convergence requirement in Assumption 3.3.4 is a natural counterpart to conditions in the discrete-time literature on co-integrating regressions. For example, replacing a sum with an integral in the discrete time setting of Vogelsang and Wagner (2014) might suggest modeling

$$X_t = X_0 + \int_0^t U_{x\tau} d\tau. \tag{3.12}$$

for some stationary process $\left\{U_{xt} \in \mathbb{R}^{d\times 1}, t \in [0,T]\right\}$. Then Assumption 3.3.4 is equivalent to an FCLT for the stationary process $\left\{U_t = (U_{0t}', U_{xt}')' \in \mathbb{R}^{d+1}, t \in [0,T]\right\}$ provided that $X_0 = o_p(T^{1/2})$. However, the form in (3.12) is not particularly desirable, and Assumption 3.3.4 is more flexible. For example, the data generating process in the non-stationary simulation environment of Section 3.4 satisfies Assumption 3.3.4. There, $\{X_t\}$ follows a two-dimensional Brownian motion and $\{U_{0,t}\}$ is a stationary Ornstein Uhlenbeck process that may not be independent of $\{X_t\}$. Alternatively to (3.12), we may view the continuous-time generalization of the setting in Vogelsang and Wagner (2014) and Hwang and Sun (2018) as requiring that, up to terms that are $o_p(T^{1/2})$, $\{X_t\}$ possesses some form of stationary increments that may be correlated with U_{0t} such that Assumption 3.3.4 holds. Viewing continuous time I(1) processes as nonstationary processes with stationary increments is adopted, for example, in Comte (1999).

In our asymptotic development, it is convenient to use the Cholesky form of $\Omega^{1/2}$ so that

$$B(\cdot) = \begin{pmatrix} B_0(\cdot) \\ B_x(\cdot) \end{pmatrix} = \begin{pmatrix} \sigma_{0\cdot x} W_0(\cdot) + \sigma_{0x}\Omega_{xx}^{-1/2} W_x(\cdot) \\ \Omega_{xx}^{1/2} W_x(\cdot) \end{pmatrix}, \tag{3.13}$$

where $\sigma_{0\cdot x}^2 = \sigma_0^2 - \sigma_{0x}\Omega_{xx}^{-1}\sigma_{x0}$ and $\Omega_{xx}^{1/2}$ is a symmetric matrix square root of Ω_{xx} such that $\Omega_{xx}^{1/2}\Omega_{xx}^{1/2} = \Omega_{xx}$.

For $j = 1, \ldots, K$, define

$$\eta_j = \int_0^1 \phi_j(r) B_x(r) dr, \ \xi_j = \int_0^1 \phi_j(r) dB_x(r),$$
$$\tilde{\nu}_j = \int_0^1 \phi_j(r) dW_0(r), \ \nu_j = \int_0^1 \phi_j(r) dB_0(r) = \sigma_{0\cdot x}\tilde{\nu}_j + \xi_j'\theta_0,$$

for $\theta_0 = \Omega_{xx}^{-1}\sigma_{x0}$ and

$$\eta = (\eta_1, \ldots, \eta_K)' \in \mathbb{R}^{K\times d}, \ \xi = (\xi_1, \ldots, \xi_K)' \in \mathbb{R}^{K\times d}, \ \zeta = (\eta, \xi) \in \mathbb{R}^{K\times 2d},$$
$$\tilde{\nu} = (\tilde{\nu}_1, \ldots, \tilde{\nu}_K)' \in \mathbb{R}^{K\times 1}, \ \nu = (\nu_1, \ldots, \nu_K)' \in \mathbb{R}^{K\times 1}.$$

Then $\nu = \xi\theta_0 + \sigma_{0\cdot x}\tilde{\nu}$.

Next, we make an assumption similar to Assumption 3.3.3.

Assumption 3.3.5 *With probability one, $\zeta'\zeta$ is of full rank $2d$.*

Let $\tilde{\Delta} x_i = (x_i - x_{i-1})/\delta$. Augmenting the discrete-time model by $\tilde{\Delta} x_i$, we obtain

$$y_i = \alpha_0 + x_i'\beta_0 + \tilde{\Delta} x_i'\theta_0 + u_{0\cdot xi},$$

where $u_{0\cdot x,i} = u_{0i} - \tilde{\Delta x}_i'\theta_0$. Using the transformed variables $\{\mathbb{W}_j^y, \mathbb{W}_j^\alpha, \mathbb{W}_j^x, \mathbb{W}_j^{\tilde{\Delta x}}, \mathbb{W}_j^{u_{0\cdot x}}\}_{j=1}^K$ defined similarly as in (3.7), we have

$$\mathbb{W}_j^y = \mathbb{W}_j^\alpha \alpha_0 + \mathbb{W}_j^x \beta_0 + \mathbb{W}_j^{\tilde{\Delta x}} \theta_0 + \mathbb{W}_j^{u_{0\cdot x}}$$

where, for example, $\mathbb{W}_j^\alpha = n^{-1/2}\sum_{i=1}^n \phi_j(i/n)$ and $\mathbb{W}_j^{u_{0\cdot x}} = n^{-1/2}\sum_{i=1}^n \phi_j(i/n)\, u_{0\cdot x,i} = \mathbb{W}_j^{u_0} - \mathbb{W}_j^{\tilde{\Delta x}}\theta_0$. Our test of $H_0 : R\beta_0 = r$ is based on estimating the above transformed and augmented regression by OLS. We call the estimator the TAOLS estimator. We outline the steps below:

1. Create the transformed variables $\{\mathbb{W}_j^y, \mathbb{W}_j^x, \mathbb{W}_j^{\tilde{\Delta x}}\}_{j=1}^K$ and stack them to form the data matrices $\mathbb{W}^y$, $\mathbb{W}^x$, and $\mathbb{W}^{\tilde{\Delta x}}$. For example, $\mathbb{W}^{\tilde{\Delta x}} = (\mathbb{W}_1^{\tilde{\Delta x}}, \ldots, \mathbb{W}_K^{\tilde{\Delta x}})' \in \mathbb{R}^{K\times d}$.

2. Regress $\mathbb{W}^y$ on $\mathbb{W}^x$ and $\mathbb{W}^{\tilde{\Delta x}}$ by OLS. Do not include an intercept. Denote the coefficients associated with $\mathbb{W}^x$ by $\hat{\beta}_{TAOLS}$, the coefficients associated with $\mathbb{W}^{\tilde{\Delta x}}$ by $\hat{\theta}_{TAOLS}$, and let $\hat{\mathbb{W}}^{u_{0\cdot x}}$ be the residual vector from this regression. Combining the matrices $\mathbb{W}^x$ and $\mathbb{W}^{\tilde{\Delta x}}$ into $\widetilde{\mathbb{W}} = (\mathbb{W}^x, \mathbb{W}^{\tilde{\Delta x}})$, we can write these objects as

$$\underset{2d\times 1}{\hat{\gamma}} \equiv \begin{pmatrix} \hat{\beta}_{TAOLS} \\ \hat{\theta}_{TAOLS} \end{pmatrix} = (\widetilde{\mathbb{W}}'\widetilde{\mathbb{W}})^{-1}\widetilde{\mathbb{W}}'\mathbb{W}^y, \;\; \hat{\mathbb{W}}^{u_{0\cdot x}} := \mathbb{W}^y - \widetilde{\mathbb{W}}\hat{\gamma}. \tag{3.14}$$

3. Calculate the test statistic

$$F_{TAOLS} = \frac{1}{\hat{\sigma}_{0\cdot x}^2}(R\hat{\beta}_{TAOLS} - r)'\left[R\left(\mathbb{W}^{x\prime}M_{\tilde{\Delta x}}\mathbb{W}^x\right)^{-1}R'\right]^{-1}(R\hat{\beta}_{TAOLS} - r)/p, \tag{3.15}$$

where $M_{\tilde{\Delta x}} = \mathbb{I}_K - \mathbb{W}^{\tilde{\Delta x}}(\mathbb{W}^{\tilde{\Delta x}\prime}\mathbb{W}^{\tilde{\Delta x}})^{-1}\mathbb{W}^{\tilde{\Delta x}\prime}$ and

$$\hat{\sigma}_{0\cdot x}^2 = \frac{1}{K}\sum_{j=1}^K (\hat{\mathbb{W}}_j^{u_{0\cdot x}})^2 = \frac{1}{K}(\hat{\mathbb{W}}^{u_{0\cdot x}})'\hat{\mathbb{W}}^{u_{0\cdot x}}. \tag{3.16}$$

These three steps are identical to the procedure in Hwang and Sun (2018) except that

$\tilde{\Delta}x_i$, instead of Δx_i, is used in the augmented regression. Such a modification serves to facilitate theoretical developments only. Since $\tilde{\Delta}x_i$ is proportional to Δx_i, the modification has no effect on the test statistic F_{TAOLS}. For practical implementation, we can follow exactly the same procedure as in Hwang and Sun (2018), utilizing Δx_i in place of $\tilde{\Delta}x_i$. There is no need to know the value of δ or its unit. We note that the test statistic in (3.15) is constructed in the same way as in the discrete-time setting.

Theorem 3.3.2 *Let Assumptions 3.2.2, 3.3.1, 3.3.4, and 3.3.5 hold. Denote $\gamma_0 = (\beta_0', \theta_0')'$ and*

$$\Upsilon_T = \begin{pmatrix} T\mathbb{I}_d & \underset{d\times d}{0} \\ \underset{d\times d}{0} & \mathbb{I}_d \end{pmatrix}.$$

(a) As $T \to \infty$ for a fixed K,

$$\left[(nT)^{-1/2}\mathbb{W}^x, \delta^{1/2}\mathbb{W}^{\tilde{\Delta}x}, \delta^{1/2}\mathbb{W}^{u_0}\right] \Rightarrow (\eta, \xi, \nu).$$

(b) As $T \to \infty$ for a fixed K,

$$\Upsilon_T(\hat{\gamma} - \gamma_0) \Rightarrow \sigma_{0\cdot x}\left(\zeta'\zeta\right)^{-1}\zeta'\tilde{\nu}.$$

In particular,

$$T(\hat{\beta}_{TAOLS} - \beta_0) \Rightarrow \sigma_{0\cdot x}\left(\eta' M_\xi \eta\right)^{-1}\eta' M_\xi \tilde{\nu} \overset{d}{=} MN\left[0, \sigma_{0\cdot x}^2\left(\eta' M_\xi \eta\right)^{-1}\right],$$

where $M_\xi = \mathbb{I}_K - \xi(\xi'\xi)^{-1}\xi'$ and "MN" stands for "mixed normal".

(c) If $K \geq 2d+1$, then, as $T \to \infty$ for a fixed K,

$$F_{TAOLS} \Rightarrow \frac{K}{K-2d} \cdot F_{p,K-2d},$$

where $F_{p,K-2d}$ is the F distribution with degrees of freedom p and $K-2d$.

Theorem 3.3.2 shows that the testing procedure of Hwang and Sun (2018) adapts to the continuous-time setting without any modification: the asymptotic F test is, therefore, robust to the sampling frequency of the data. From an applied point of view, we do not have to be concerned about whether we have high-frequency data with a shrinking sampling interval (i.e., $\delta \to 0$) or discrete-time data with a fixed sampling interval (e.g., $\delta = 1$). This gives us much practical convenience.

To implement the above F test, we follow the procedure below, which is similar to that in the exogenous case.

1. Estimate the model $y_i = \alpha_0 + x_i'\beta_0 + u_{0i}$ by OLS to obtain the residual

$$\hat{u}_{0i} = y_i - \hat{\alpha}_{OLS} - x_i'\hat{\beta}_{OLS}.$$

2. On the basis of $\{\hat{u}_{0i}\}$, use the series method to estimate the long run variance of $\{u_{0i}\}$, computing the AR(1) data-driven $\hat{K}_D$ using the formula in (3.5).

3. Let $\hat{K}^* = \max(\hat{K}_D, 2d+3)$ and use $\hat{K}^*$ to construct the transformed and augmented regression.

4. Compute the F test statistic in the TAOLS regression. Perform the asymptotic F test using $\frac{\hat{K}^*}{\hat{K}^*-2d} \cdot F_{p,\hat{K}^*-2d}$ as the reference distribution.

3.4 Simulation Evidence

In this section, we conduct simulations to evaluate the finite-sample size and power properties of the proposed F tests. For the stationary setting, we consider the model

$$Y_t = \beta_{01} + X_t\beta_{02} + U_t, \;\; 0 \le t \le T,$$

with $\beta_{01} = 0$ and $\beta_{02} = 1$. We test $H_0 : (\beta_{01}, \beta_{02})' = (0,1)'$ versus $H_1 : (\beta_{01}, \beta_{02})' \neq (0,1)'$. (X_t) and (U_t) are chosen as stationary Ornstein-Uhlenbeck (OU) processes described by

$$dX_t = -\kappa_x X_t dt + \sigma_x dV_t \text{ and } dU_t = -\kappa_u U_t dt + \sigma_u dW_t,$$

where $(\kappa_x, \sigma_x) = (0.1020, 1.5514)$, $(\kappa_u, \sigma_u) = (6.9011, 2.7566)$, and (V_t) and (W_t) are independent standard Brownian motions. The parameter values of the OU processes are obtained from Chang et al. (2021), who estimate (κ_x, σ_x) by fitting an OU process to 3-month T-bill rates from 1971 to 2016 and estimate (κ_u, σ_u) by fitting an OU process to the residuals obtained by regressing 3-month eurodollar rates on these T-bill rates. As an alternative to an OU explanatory variable process, we also consider the process $X_t = C_t - \mu_c$ where

$$dC_t = \kappa_x (\mu_c - C_t) dt + \sigma_x \sqrt{C_t} dV_t,$$

and V_t is again standard Brownian motion. This corresponds to Feller's Square Root (SR) process. In this setting, we keep the OU process $\{U_t\}$ as described above (again with $\{W_t\}$ independent of V_t) and $(\mu_c, \kappa_x, \sigma_x) = (4.8196, 0.1794, 0.9367)$ where these parameters come from fitting the SR process to 3-month T-bill rates from 1971 to 2016.

In the nonstationary setting, we consider the model

$$Y_t = \alpha_0 + X_{1,t}\beta_{01} + X_{2,t}\beta_{02} + U_{0t}, \quad 0 \leq t \leq T,$$

with $\alpha_0 = 0, \beta_{01} = 1, \beta_{02} = 1$. We test $H_0 : (\beta_{01}, \beta_{02})' = (1,1)'$ versus $H_1 : (\beta_{01}, \beta_{02})' \neq (1,1)'$. In this setting, we model $(X_{j,t})$, $j \in \{1,2\}$, as Brownian motions and (U_{0t}) as a stationary OU process. In particular, for $j \in \{1,2\}$, we have

$$dX_{j,t} = \sigma_j dZ_{j,t} \text{ and } dU_{0t} = -\kappa_u U_{0t} dt + \sigma_u dZ_{3,t},$$

where $\sigma_1 = \sigma_2 = 0.0998$, $(\kappa_u, \sigma_u) = (1.5717, 0.0097)$, and

$$\begin{pmatrix} Z_{1,t} \\ Z_{2,t} \\ Z_{3,t} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \varphi & \sqrt{1-\varphi^2} & 0 \\ \varphi & \frac{\varphi-\varphi^2}{\sqrt{1-\varphi^2}} & \sqrt{1-\left(\varphi^2 + \frac{(\varphi-\varphi^2)^2}{1-\varphi^2}\right)} \end{pmatrix} \begin{pmatrix} W_{1,t} \\ W_{2,t} \\ W_{3,t} \end{pmatrix}.$$

Here $W_{1,t}$, $W_{2,t}$, and $W_{3,t}$ are independent standard Brownian motions and $\varphi \geq 0$. In this setup, each $(Z_{j,t}), j \in \{1,2,3\}$, is a standard Brownian motion and $Corr(Z_{k,t}, Z_{\ell,t}) = \varphi$ when $k \neq \ell$. The parameter values here also originate from Chang et al. (2021); (σ_1) comes from fitting a Brownian motion process to log US/UK exchange rate spot price data from 1979 to 2017. (κ_u, σ_u) are estimated by fitting an OU process to the residuals from regressing log US/UK exchange rate forward prices on the log US/UK exchange rate spot prices. We consider both $\varphi = 0$ (the exogeneous case) and $\varphi = 0.75$ (the endogenous case).

In addition to the baseline values of κ_x and κ_u, we also multiply κ_x and κ_u by 4 and $1/4$, allowing for variation in the mean reversion parameters of the stationary elements of the simulations. As the mean reversion parameter gets closer to zero, the stationary OU (or SR) process becomes more persistent and in the OU case behaves more like a nonstationary Brownian motion.

In both the stationary and nonstationary settings, we consider $T = 30$ and $T = 60$. The stochastic processes are generated using the transition densities of Brownian motion, OU, and SR processes except in the nonstationary case when $\varphi = 0.75$. In this case, transition densities are used to generate all processes except that U_t is constructed via Euler's method once $Z_{3,t}$ is generated. Discrete samples are collected at various frequencies between $\delta = 1/252$ and $\delta = 1/4$. In each scenario, we replicate the simulation 5000 times.

To implement the testing procedures described in the earlier sections, we utilize the sine and cosine basis functions given in (3.3) and choose K via the data-driven procedures described in Sections 3.2 and 3.3. In our figures described below, results corresponding to these tests are denoted “Series F”, and there are different figures for the stationary and nonstationary settings. As K increases, in both the stationary and nonstationary settings, the limiting distributions of the test statistics approach the scaled chi-squared distribution χ_p^2/p. The scaled chi-squared approximation can also be obtained by letting $K \to \infty, \delta \to 0$ and $T \to \infty$ jointly[5]. Utilizing the critical values from this distribution with our test statistics, we denote the resulting results by “Series Chi2.” In the figures for the nonstationary setting, “Series F” and “Series Chi2” are reserved for the procedure outlined in Subsection 3.3.2 that can accommodate endogeneity. These labels are replaced by “S-EXO F” and “S-EXO Chi2” , respectively, for the procedures designed where $\{U_t\}$ is assumed exogenous described in Subsection 3.3.1.

To compare the F tests with some existing tests, we carry out the kernel-based tests of Chang et al. (2021). For their tests, we employ the quadratic spectral (QS) kernel and utilize Andrews (1991)’s bandwidth selection procedure, which is among the best performers in the simulations in Chang et al. (2021). In our figures, the results corresponding to the QS kernel are denoted “Kernel Chi2.” To include the fixed-b version of their tests, we note that the test statistics of Chang et al. (2021) in the stationary setting and the nonstationary setting with exogeneous regressors, without any change in form, have fixed-b counterparts in the discrete-time settings of Kiefer and Vogelsang (2005) and Jin et al. (2006), respectively. Utilizing arguments similar to what we present here and in Vogelsang and Wagner (2014), it is not difficult to ascertain that the limiting distributions identified in these papers are also applicable in our simulation set up with exogenous regressors. In the cointegrating regression with endogenous regressors, the fixed-b asymptotics of Jin et al. (2006) is not applicable to the test statistic of Chang et al. (2021), as it does not account for endogeneity. To use the fixed-b asymptotics of Vogelsang and Wagner (2014), which accounts for endogeneity, we have to run a different set of regressions and alter the test statistic. This would require further theoretical development and is not considered in our simulations. The tests utilizing the fixed-b approximations of Kiefer and Vogelsang (2005) and Jin et al. (2006) for the test statistics in Chang et al. (2021) are denoted by “Kernel fixed-b” in

[5] The scaling factor of $1/p$ arises because the test statistics are scaled by p.

our figures.

3.4.1 Size study

Figures 2 – 5 display the empirical sizes (i.e., the null rejection probabilities) in the different simulation scenarios.

Figures 2 and 3 show that in the stationary setting, the series-based F test exhibits less size distortion than all chi-squared tests under consideration. The improvement in the size accuracy of the F test over the chi-square tests is more visible when the underlying OU or SR processes have smaller mean reversion parameters κ_x and κ_u and thus become more persistent. This is consistent with the literature on HAR inference in the discrete-time setting. See, for example, Sun (2013), Sun (2014b), Sun et al. (2008), and Kiefer and Vogelsang (2005) for simulation evidence and theoretical developments. The F test performs similarly to the fixed-b version of the test in Chang et al. (2021) adapted from Kiefer and Vogelsang (2005). This is expected, because both types of tests utilize nonparametric LRV estimators, and both are based on fixed-smoothing asymptotic approximations. The advantage of the series-based F test is that it is more convenient to use, as critical values are readily available from statistical tables and standard programming environments. There is no need to simulate a nonstandard fixed-smoothing asymptotic distribution, an unavoidable and formidable task if we use a kernel-based fixed-smoothing test. We note in passing that all chi-squared tests have similar performances, regardless of whether series-based or kernel-based LRV estimators are used. This provides further simulation evidence that the type of LRV estimators used does not matter much. What matters more is the reference distribution used in a testing procedure.

In the nonstationary setting with exogenous regressors, the performance of the F tests relative to the fixed-b version of the test in Chang et al. (2021) adapted from Jin et al. (2006) and the chi-squared tests is qualitatively similar to that in the stationary setting. In particular, the F tests and the fixed-b test achieve more or less the same size control. However, the fixed-b tests in this setting aren't developed fully for the continuous-time setting. The validity of the fixed-b test relies not only on the exogeneity of the regressors but also crucially on the premise that the limiting process (X°) is a Brownian motion process. Similarly, the F-test of Subsection 3.3.2 designed for potential endogeneity also relies on a Brownian motion limiting process for its validity. While this does not cause problems in our simulation setting where the premise holds, the fixed-b asymptotic distribution and that associated with the F-test in Subsection 3.3.2 are, in general, functionals of (X°), which may contain additional nuisance parameters beyond its scale. A benefit of our approach in Subsection 3.3.1 is that the conditioning argument in the proof of Theorem 3.3.1 bypasses reliance on the distributional form of (X°). Such a conditioning argument does not go through if we use a kernel LRV estimator.

In the nonstationary setting with endogenous regressors, to the best of our knowledge, the F test in Subsection 3.3.2 appears to be the only asymptotically valid test in the literature. Unsurprisingly, it exhibits better size properties than the alternative tests from the pre-existing literature, including the fixed-b version of the test in Chang et al. (2021). While the F-test of Subsection 3.3.1 which assumes the error process $\{U_t\}$ is exogenous appears to maintain competitiveness against the F test of Subsection 3.3.2, this unfortunately is an artifact of the particular DGPs in our simulation setting. In this simulation environment, it can be shown that the limiting distribution of the exogeneity-based test is a noncentral F distribution that depends on nuisance parameters. The F distribution used happens to be relatively close to the finite sample distribution but will result in a poor approximation in general. We note that the presence of the endogeneity bias can lead to a large size distortion, especially when the chi-square approximation is used. For example, when $\varphi = 0.75$, $T = 30$, and κ_u is 1/4 of the baseline value, the null rejection probability of the 5% chi-squared test of Chang et al. (2021) can be as high as 60%.

Figures 2 – 5 further show that the size properties of all tests are not sensitive to the sampling interval δ, and all tests become more accurate when T increases. This is consistent with our theoretical results that the effective sample size is T and is unrelated to δ. Intuitively, for a given time span T, as δ decreases, the number of sampled observations n increases, but at the same time, the sampled observations become more persistent. These two effects offset each other, leading to an effective sample size of T.

3.4.2 Power study

Figures 6 – 8 investigate the empirical power properties of the test procedures in finite samples; the power is size-adjusted. To evaluate the power of the tests, we use the baseline designs. When generating the data, each of the parameters being tested is multiplied by $1-\psi$ for a range of $\psi \in [0,1]$. To keep the visualization simple, we focus only on the frequencies $\delta = 1/252$ and $\delta = 1/4$. As there are only two different test statistics, ours and that in Chang et al. (2021) and the power is size-adjusted, there are only two different sized-adjusted power curves. The reported figures only display the comparison for the series-based approach in Sections 3.2 and 3.3 and the kernel-based approach in Chang et al. (2021). In the figures, the higher frequency $\delta = 1/252$ is denoted "h", and the lower frequency $\delta = 1/4$ is denoted "l."

Figures 6 and 7 show that, in the stationary setting, all tests have almost indistinguishable power curves. In the nonstationary setting with exogenous regressors, the series-based tests have competitive power relative to the kernel-based tests, although when $T = 30$ the former are slightly less powerful most noticeably in the procedure that allows for endogeneity. This could be explained by the MSE-optimality of the QS kernel among the second-order positive-definite kernels. In the nonstationary setting with endogenous regressors, the comparison is not as meaningful, as the tests of Chang et al. (2021) have significant size distortion. Nevertheless, the series-based tests still have competitive power, especially when $T = 60$. When $T = 30$, the series-based tests are somewhat less powerful.

Figures 6 and 8 also show that the power properties of all tests are not sensitive to the choice of δ. In each scenario, the power curves for $\delta = 1/252$ and $\delta = 1/4$ are virtually identical. This echoes the finding that the size properties are not sensitive to δ. In each scenario, all tests become more powerful when T is larger, reflecting that it is the time span T, not the number of observations n, that is the effective sample size.

3.5 Empirical Application

Here we examine the series-based F test in an application to interest rate data that are available at multiple sampling frequencies. In particular, we revisit an application appearing in Chang et al. (2021), which focuses on characterizing the co-movements of interest rates among securities with different times to maturity. As discussed in Chang et al. (2021), the ability of the U.S. Federal Reserve System (FED) to influence long-term interest rates via the short-term Federal Funds Rate (FFR) was challenged during the Global Financial Crisis (GFC) of 2008 when the zero lower bound for the FFR was reached. This partially motivated the FED's adoption of non-conventional policies such as quantitative easing. To investigate the dynamics between short and long rates within their linear hypothesis testing methodology, Chang et al. (2021) test for "parallel shifts" among securities with varying maturities. Here, "parallel shifts" refers to changes in the yields of securities with different maturities tending to be of the same size and direction. Chang et al. (2021) regress 10-year U.S. Treasury bond (T-bond) yields on 3-month Treasury bill (T-bill) yields and consider data before and after the GFC separately. The existence of "parallel shifts" would imply a slope coefficient near one, and Chang et al. (2021) find that, prior to the GFC, there is no strong evidence against the null hypothesis of a unit slope coefficient. This is consistent with the view that the FED was able to successfully influence long rates via short rate policies prior to the GFC.

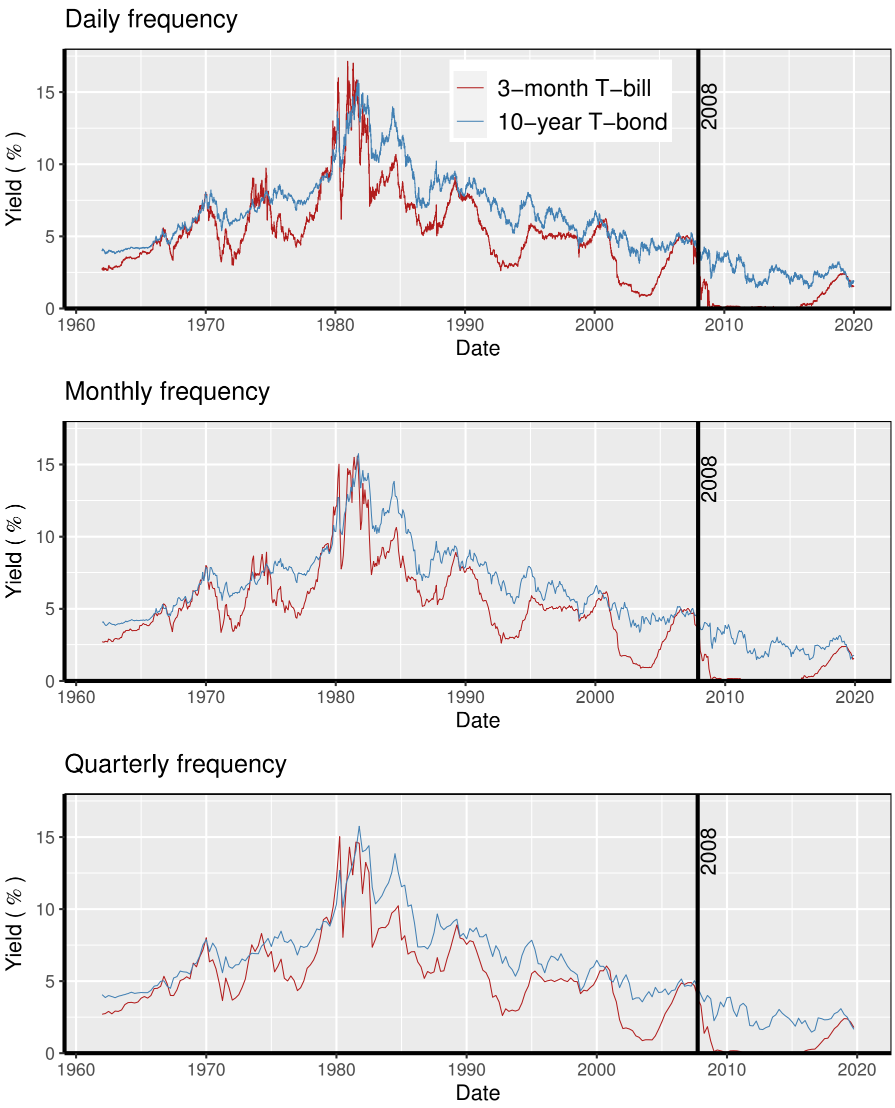
This regression setting, detailed below, is useful for evaluating our testing procedure because it is simple and allows for the consideration of several hypothesis tests of varying theoretical credibility. For example, the additional null hypothesis that the intercept coefficient is zero states that, on average, the yield spread is zero. If the yields of U.S. government securities of different duration differ based on compensation for interest rate risks and the expectations of future interest rates, we may expect to reject this hypothesis. Additionally, as the setting has been analyzed in Chang et al. (2021), we may contrast our methodology and results with theirs. We find that the conclusions stemming from the F tests are largely in line with those from the testing procedures of Chang et al. (2021). We observe, however, that the F test for one hypothesis test of interest produces a less ambiguous result at the daily sampling frequency and also bypasses a subjective modeling decision that can inflate one of the test statistics analyzed in Chang et al. (2021).

The continuous-time regression of interest is given by

$$Y_t = \alpha + X_t\beta + U_t,$$

where Y_t is the yield (in percent) of 10-year T-bonds at time t and X_t is the yield of 3-month T-bills. We observe $\{X_{i\delta}\}_{i=1}^{n}$ and $\{Y_{i\delta}\}_{i=1}^{n}$ at three fixed sampling interval lengths, δ, corresponding to daily, monthly, and quarterly frequencies. The number of observations n varies with δ as each sample is derived from a fixed time span, but we do not complicate the notation here. Recall, additionally, that the F test would be valid if applied to discrete time series under the standard discrete-time assumptions. The two yield series of different maturities are available from the Federal Reserve Economic Data (FRED) of the St. Louis FED. As in Chang et al. (2021), we consider three null hypotheses independently of one another and we consider two different sample windows. All hypothesis tests are performed twice, once utilizing data from each sample window separately. The null hypotheses are $H_0^{\alpha} : \alpha = 0$, $H_0^{\beta} : \beta = 1$, and $H_0^{\alpha,\beta} : \alpha = 0$ and $\beta = 1$ jointly. The first sample window includes data from 1962 to 2007 while the second contains observations from 2008 to 2019.

The two interest rate series plotted at the various sampling frequencies are presented in Figure 1. In Table 1, we present the test statistics associated with the various null hypotheses for each sample window. Test statistics titled "Series-F" refer to the F test described in Section 3.2 designed around the stationary regression setting. Those under the header "Kernel-χ^2" are performed utilizing the kernel-based χ^2 test of Chang et al. (2021) which they refer to as the H-test. The χ^2 tests (i.e., H-tests) in Figure 1 are calculated using the Andrews (1991) bandwidth procedure which is "high-frequency-compatible" as discussed in Chang et al. (2021) and utilizing the QS kernel. Rejection of a null hypothesis at the 5% level is indicated by “*” and rejection at the 1% level is indicated by “**”. P-values are included in brackets for testing the null of “parallel shifts” H_0^β.


Daily frequency

Monthly frequency

Quarterly frequency

Figure 1. 10-year Treasury bond and 3-month Treasury bill yields at the sampling frequencies analyzed. A line at the beginning of 2008 demarcates the two sample windows.

We can see from Table 1 that the results of the F tests are stable across all sampling frequency choices. This is consistent with the theory developed earlier in the paper, namely that the tests are valid for high-frequency observations over a long span and have direct counterparts that are valid and familiar in the discrete-time setting when the sampling frequency is lower. For the F tests, the statistical conclusions reached for each null hypothesis and sampling window remain the same for each sampling frequency: all null hypotheses are rejected at the 1% level except that we are unable to reject the "parallel shifts" hypothesis H_0^{β} at even the 5% significance level in any frequency using data prior to the GFC. This evidence is consistent with the view that the FED was able to control long rates via short-term policy rates prior to the GFC. Additionally, there is evidence against the hypothesis of a zero average yield spread (H_0^{α}, which is included in $H_0^{\alpha,\beta}$) before and after the GFC of 2008. This is consistent with the stylized fact that the yield curve tends to be upward sloping. The results and conclusions of the F tests are thus in agreement with the findings of Chang et al. (2021) where χ^2-based tests with "high-frequency compatible" bandwidths are utilized. Note that their findings are mirrored by those for the kernel-based χ^2 tests reported in Figure 1 which are computed according to their methodology. In contrast, Chang et al. (2021) show that in this regression setting, tests that are not robust to the sampling frequency or utilize a bandwidth choice that is not "high-frequency compatible" will reject H_0^{β} at the daily frequency.

Lastly, we discuss some differences between the F test and the kernel-based χ^2 test of Chang et al. (2021) in this application that may be indicative of the benefits of the F test. First, note that for the kernel-based χ^2 test using pre-GFC observations at the daily sampling frequency, the test statistic surpasses the critical value for a 5% test but not that of a 1% test. Chang et al. (2021) choose to view this as failing to reject the null hypothesis, requiring that the test statistic surpass the 1% critical value to take a more conservative stance. To this end, they note that the nominal size may understate the empirical rejection probability as observed in their (and our) simulations. On the other hand, the F test statistic here fails to surpass the critical value for a 5% test, corresponding to a p-value of 0.0787. As seen in our simulations and discussed in relation to the fixed-smoothing literature in Subsection 3.4.1, the F test can result in tests with more accurate size. This example may be a case where some ambiguity regarding test significance is avoided.

Table 1. Test statistics computed with observations collected at different sampling frequencies. Brackets contain p-values. Rejection of a null hypothesis at the 5% level is indicated by "*" and rejection at the 1% level is indicated by "**". p-values are included in brackets for testing the null of "parallel shifts" H_0^{β} based on the pre-GFC observations.

Sample: 1962-2007						
Sampling Freq.	Daily		Monthly		Quarterly	
Test Stat.	Series-F	Kernel-χ^2	Series-F	Kernel-χ^2	Series-F	Kernel-χ^2
H_0^{α}	18.73**	22.10**	18.77**	20.74**	19.96**	19.93**
H_0^{β}	3.69	4.30*	3.42	3.67	3.07	3.15
	[0.0787]	[0.03801*]	[0.0874]	[0.0553]	[0.1000]	[0.0760]
$H_0^{\alpha,\beta}$	17.75**	38.26**	18.97**	38.95**	21.26**	41.43**

Sample: 2008-2019						
Sampling Freq.	Daily		Monthly		Quarterly	
Test Stat.	Series-F	Kernel-χ^2	Series-F	Kernel-χ^2	Series-F	Kernel-χ^2
H_0^{α}	87.19**	106.62**	87.68**	107.52**	129.07**	124.77**
H_0^{β}	81.44**	32.29**	81.59**	27.48**	37.94**	20.38**
$H_0^{\alpha,\beta}$	46.25**	113.73**	48.11**	116.19**	65.96**	127.73**

Another point of interest for the F test in this example is as follows. Some of the test statistics considered in Chang et al. (2021) may require/allow the researcher to determine a continuous-time modeling parameter that could influence the test statistic's magnitude. Such a test statistic utilizes a (kernel-based) LRV estimator that, when utilizing the discrete-time counterpart LRV estimator, requires a "high-frequency compatible" bandwidth parameter b_n in order to produce a valid test. One choice they consider is their continuous-time rule of thumb (CRT). This is given by

$$b_n = cn^a/\delta^{1-a},$$

where $c > 0$ and $0 < a < 1$. In contrast to discrete-time rules of thumb for kernel-based LRV bandwidth parameters, there is now a division by δ^{1-a}. However, δ depends on the unit of time that T is measured in, which may be subjective. Suppose we set $c = 2.3019$ and $a = 1/5$ and wish to test H_0^β with daily observations between 1962 and 2007. These choices for a and c correspond to a guideline in Andrews (1991) for the QS kernel in a discrete-time setting when considering an AR(1) process with coefficient 0.5. (The observation below also holds with similar test statistics and p-values if we choose the alternative discrete-time rule of thumb choices $c = 3/4$ and $a = 1/3$, suggested in the undergraduate textbook Stock and Watson (2019); see equation (16.17) there). If we assume T is measured in years, i.e., $T = 46$ years between 1962 and 2007, then $\delta = 1/252$ for about 252 trading days in a year. Alternatively, suppose we think that T should be measured in months so that $T = 552$ months. Then we may set $\delta = 1/21$ for roughly 21 trading days in a month. As we see below, this distinction changes the test conclusion.

Table 2. Test statistics computed with observations collected at a daily sampling frequency during 1962-2007 for the "parallel shifts" hypothesis H_0^β. In addition to the test statistics from earlier, additional kernel-based χ^2 test statistics of Chang et al. (2021) are presented when computed with the CRT using $\delta = 1/252$ and $\delta = 1/21$. Rejection of a null hypothesis at the 5% level is indicated by "*" and rejection at the 1% level is indicated by "**". p-values are included in brackets.

Sample: 1962-2007, Daily Frequency				
Stat.	Series-F	Kernel-χ^2-AD	Kernel-χ^2-CRT, $\delta = 1/252$	Kernel-χ^2-CRT, $\delta = 1/21$
H_0^β	3.69 [0.0787]	4.30* [0.0380]	4.46* [0.0347]	10.03** [0.0015**]

Table 2 contains the test statistics computed from daily observations between 1962 and 2007 for the null hypothesis H_0^β. In addition to the test statistics considered earlier, it includes two alternative calculations for the kernel-based χ^2 test statistic, denoted by "Kernel-χ^2-CRT, $\delta = 1/252$" and "Kernel-χ^2-CRT, $\delta = 1/21$." These correspond to the choice of δ described above. The corresponding test statistics from Table 1 are also included. The kernel-based χ^2 test of Chang et al. (2021) reported earlier in Table 1 that is calculated utilizing the procedure of

Andrews (1991) is now denoted "Kernel-χ^2-AD." Note that, like the F test statistic, this version of the test statistic does not feature a direct reliance on a user inputted δ. From Table 2, we see that changing δ from $1/252$ to $1/21$ increases the CRT-based test statistic to surpass the critical value for a 1% test. If δ is chosen too large, we get a bandwidth that is too small for a continuous-time process that varies slowly at higher frequency observations. The effect is similar to using a "high-frequency incompatible" bandwidth, a setting explored in Chang et al. (2021) that leads to spurious tests with divergent test statistics. This example suggests that tests which do not rely on a user choosing δ, such as the F test or the test of Chang et al. (2021) that utilizes the Andrews (1991) bandwidth procedure, may be more robust against debatable modeling decisions that could impact statistical significance. In addition to potential size-accuracy benefits, the F test adds to the available tests with this feature, and only one such test is discussed in Chang et al. (2021).

3.6 Conclusion

This paper provides a simple approach to linear hypothesis testing that is robust to the potential continuity of the underlying data generating processes. The test procedures demonstrate reduced size distortion in finite samples relative to existing approaches and can accommodate endogeneity in cointegration-type regressions. From a practical point of view, the tests have several desirable characteristics. Their direct correspondence to analogous discrete-time procedures clears the practitioner from modeling choices that could influence test results. Additionally, the limiting distributions do not need any complicated simulations to derive critical values as some discrete-time fixed-b approaches require; the tests rely only on standard F-distributions. In the cointegrating regression setting with exogeneous regressors, more accurate tests are delivered while maintaining greater generality with regard to the limiting behavior of the regressor process. Lastly, in the working paper version (Pellatt and Sun (2022)) of this paper, we have shown that our asymptotic F theory remains valid in the presence of additive measurement noises in the regressor error.

Chapter 3, in full, is a reprint of the material as it appears in Asymptotic F Test in Regressions With Observations Collected at High Frequency Over Long Span 2022. Pellatt, Daniel F.; Sun, Yixiao, Journal of Econometrics, 2022. Minor adjustments around the referencing and title of an appendix have been made to integrate the format with that of this dissertation. The dissertation author is a primary author of this material.

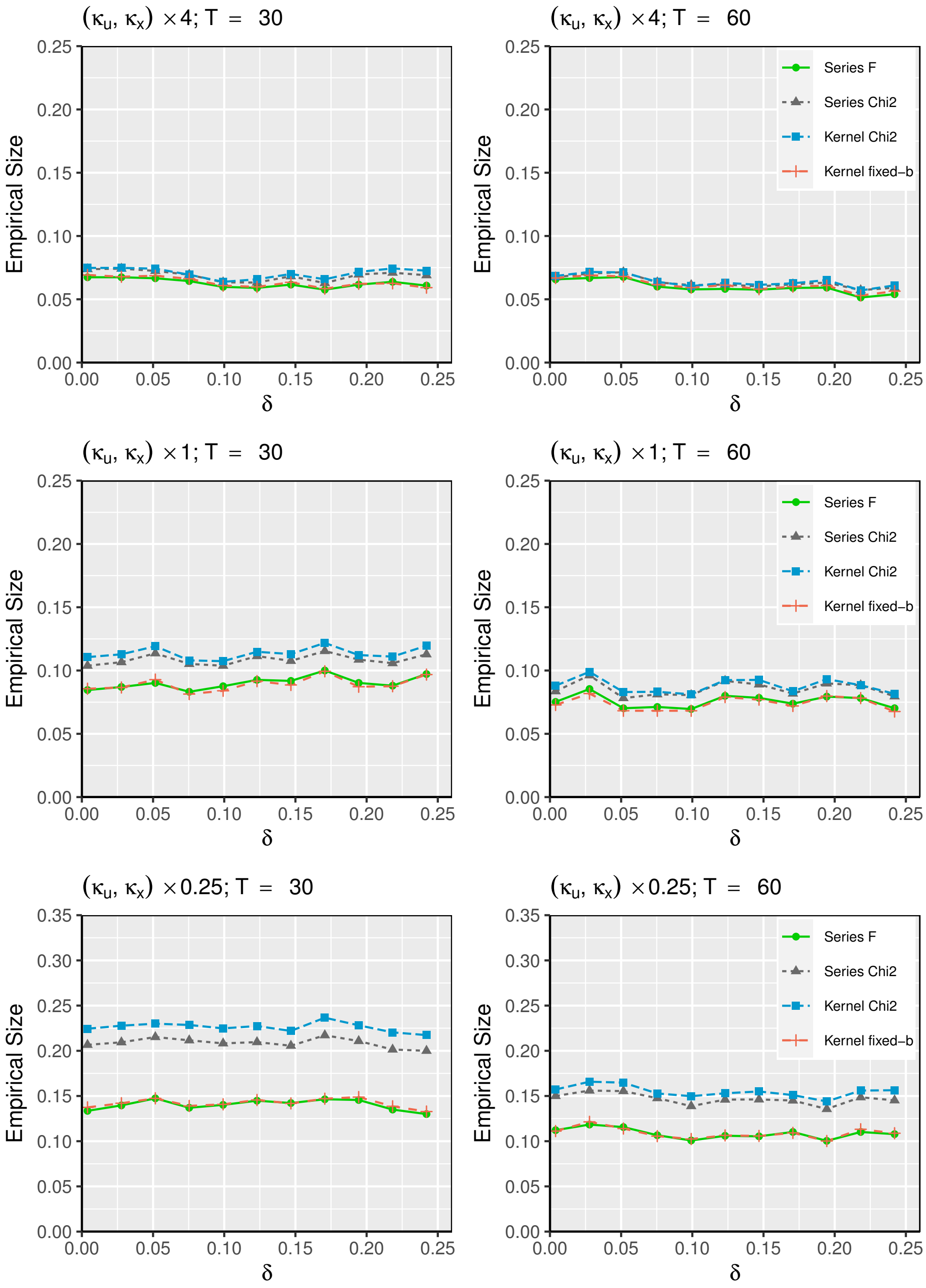


Figure 2. Empirical sizes in the stationary simulation setting when X_t follows an OU process and (κ_u, κ_x) are multiplied by factors of 4, 1 and $1/4$.

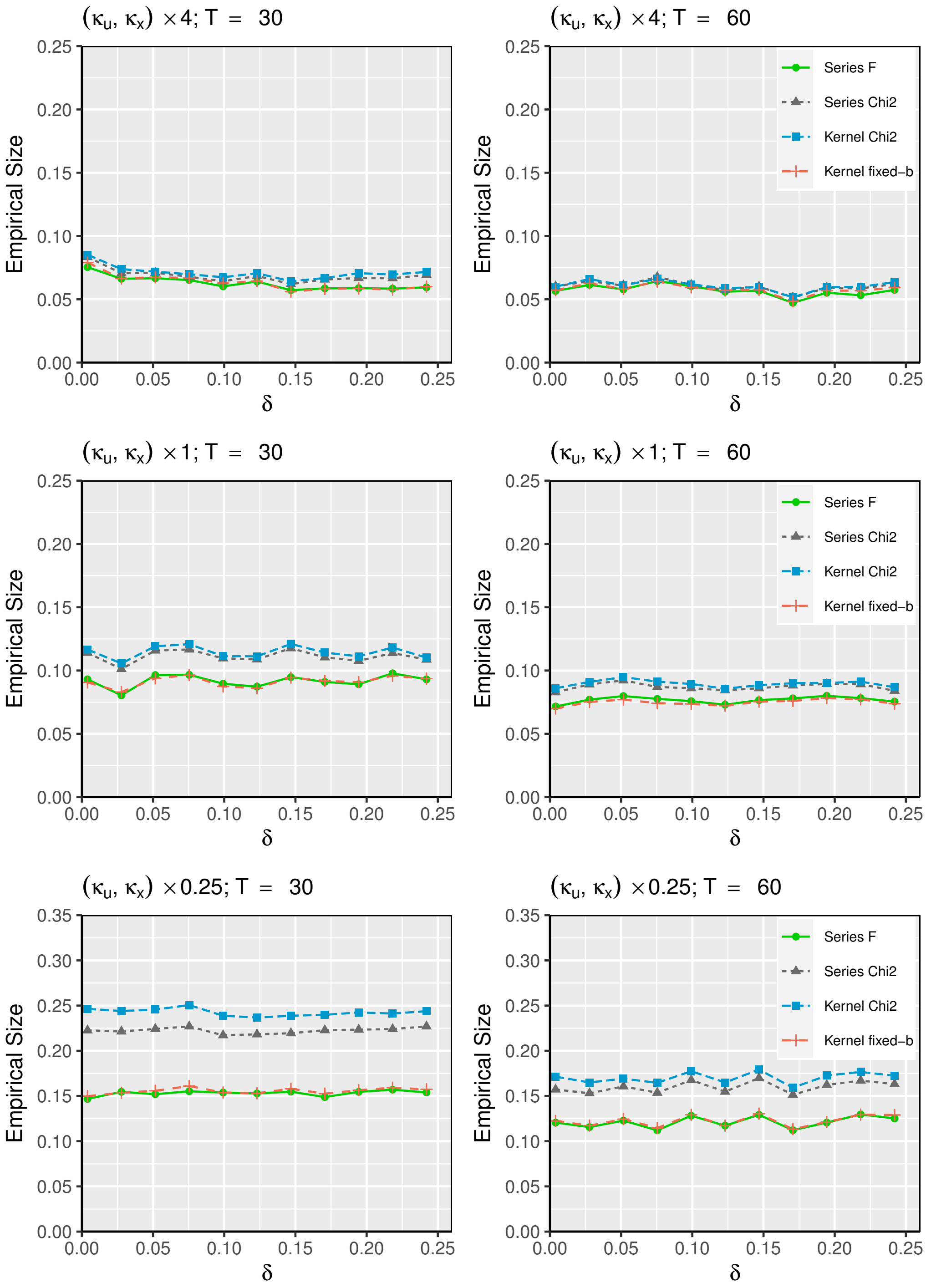


Figure 3. Empirical sizes in the stationary simulation setting when X_t follows an SR process and (κ_u, κ_x) are multiplied by factors of 4, 1, and $1/4$.

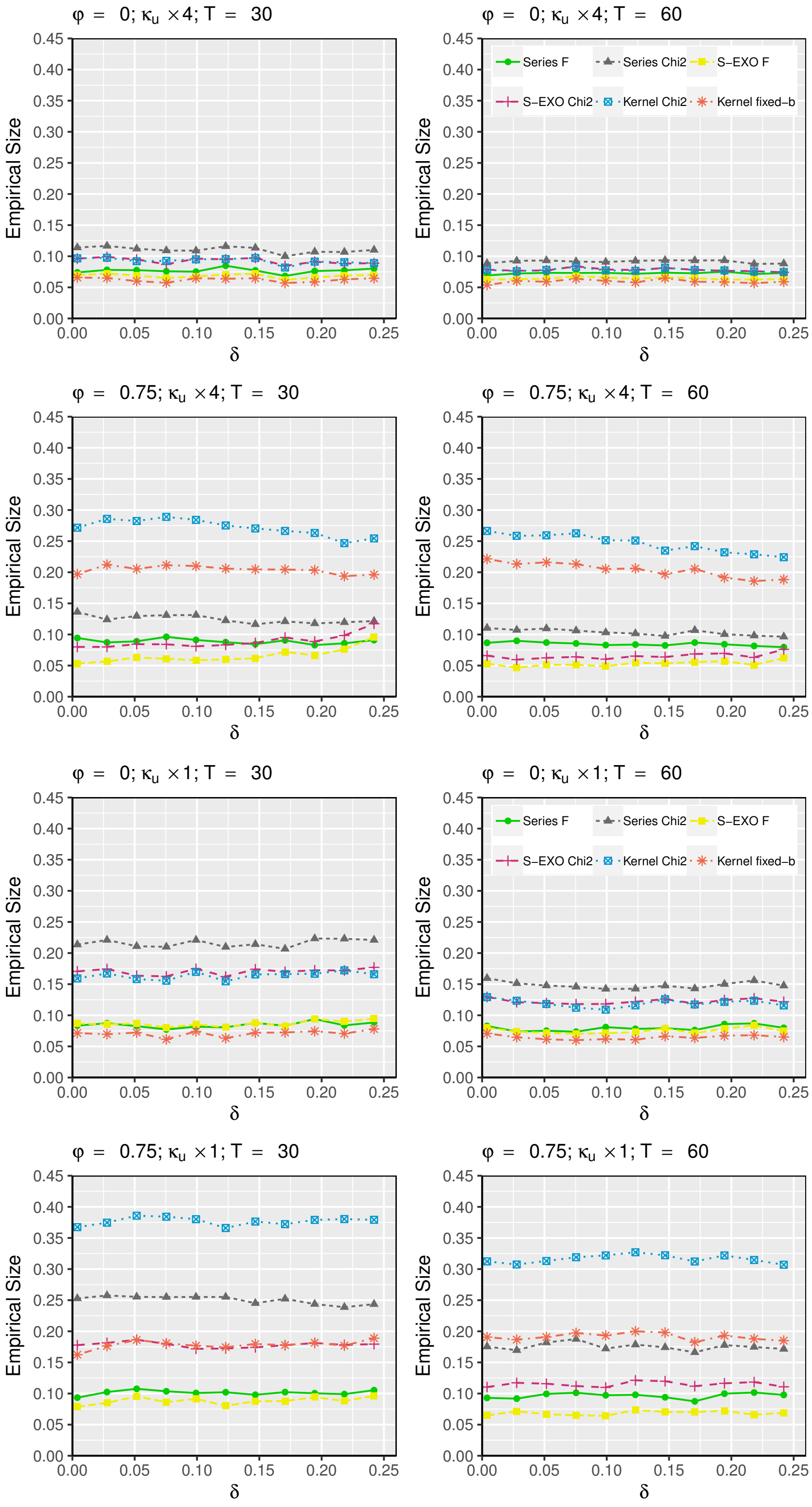


Figure 4. Empirical sizes in the nonstationary simulation setting when κ_u is multiplied by factors of 4 and 1.

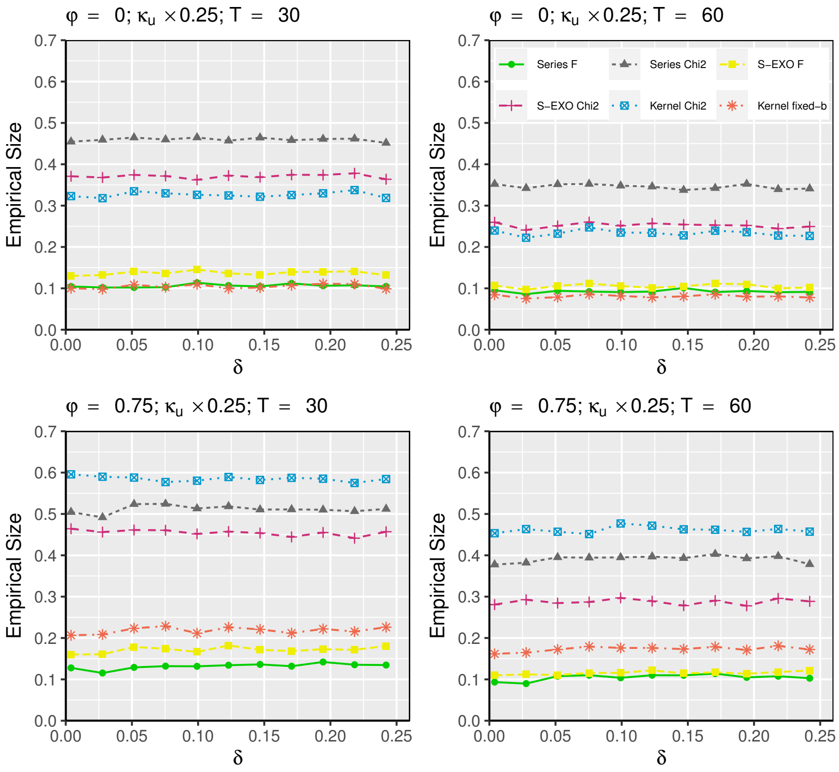


Figure 5. Empirical sizes in the nonstationary simulation setting when κ_u is multiplied by $1/4$.

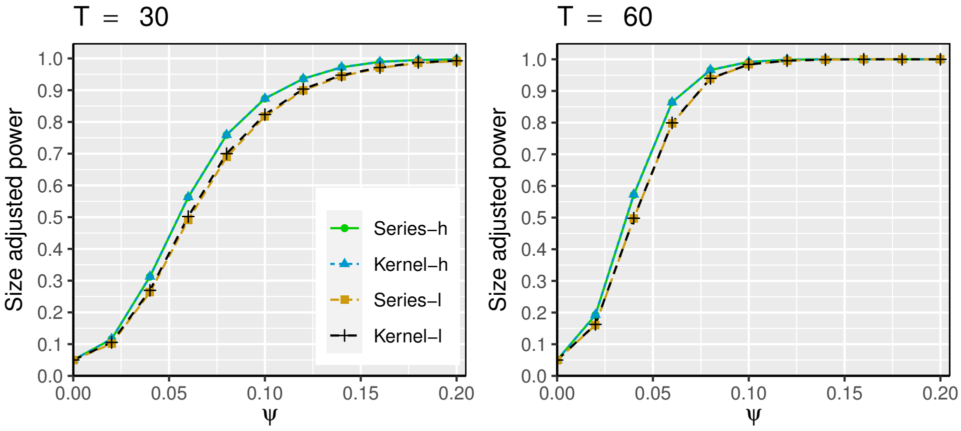


Figure 6. Size-adjusted powers in the stationary setting when X_t is distributed according to the OU process described in Section 3.4

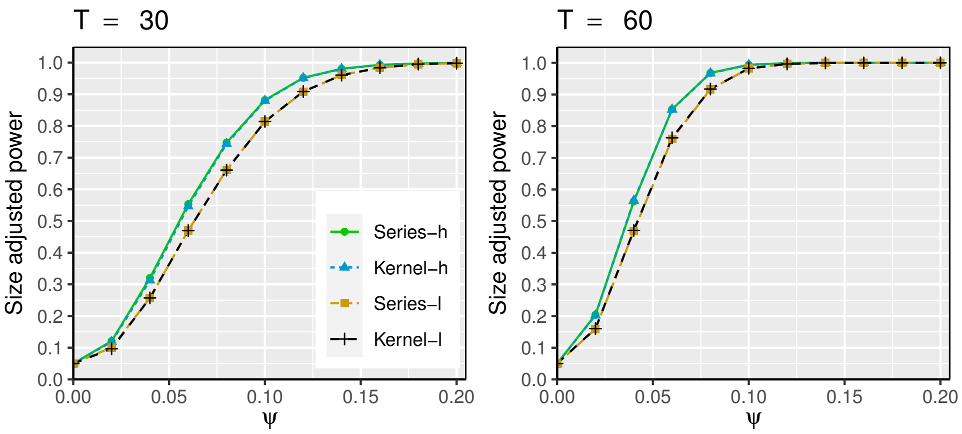


Figure 7. Size-adjusted powers in the stationary setting when X_t is distributed according to the SR process described in Section 3.4

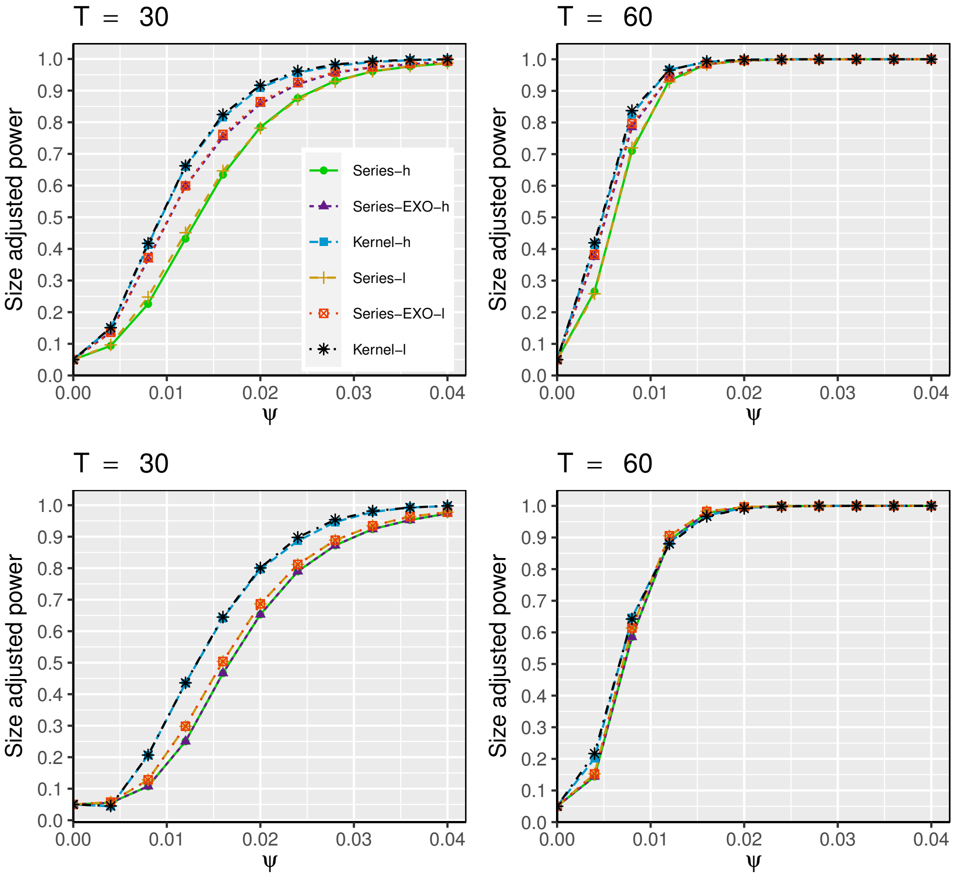


Figure 8. Size-adjusted powers in the nonstationary setting. In the upper row, the explanatory variables are exogenous ($\varphi = 0$). In the lower row the explanatory variables are endogenous ($\varphi = 0.75$).

Appendices

3.A Appendix of Proofs for Chapter 3

Proof of Lemma 3.2.1. We start by writing

$$\frac{1}{T}\int_0^T \phi_j\left(\frac{t}{T}\right)Z_t dt = \frac{1}{T}\sum_{i=1}^n \int_{(i-1)\delta}^{i\delta} \phi_j\left(\frac{t}{T}\right)Z_t dt, \tag{3.17}$$

and

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \phi_j\left(\frac{i}{n}\right)z_i &= \frac{1}{T}\sum_{i=1}^n \delta\phi_j\left(\frac{i-1}{n}\right)Z_{(i-1)\delta} + \frac{\delta}{T}\left[\phi_j\left(1\right)Z_T - \phi_j\left(0\right)Z_0\right]\\ &= \frac{1}{T}\sum_{i=1}^n \delta\phi_j\left(\frac{i-1}{n}\right)Z_{(i-1)\delta} + O_p\left(\frac{\delta}{T}\sup_{t\in[0,T]}\|Z_t\|\right).\end{aligned}$$

So,

$$\begin{aligned}&\frac{1}{T}\int_0^T \phi_j\left(\frac{t}{T}\right)Z_t dt - \frac{1}{n}\sum_{i=1}^n \phi_j\left(\frac{i}{n}\right)z_i\\ &= \frac{1}{T}\sum_{i=1}^n \int_{(i-1)\delta}^{i\delta}\left[\phi_j\left(\frac{t}{T}\right)Z_t - \phi\left(\frac{i-1}{n}\right)Z_{(i-1)\delta}\right]dt + O_p\left(\frac{\delta}{T}\sup_{t\in[0,T]}\|Z_t\|\right)\\ &= \frac{1}{T}\sum_{i=1}^n \int_{(i-1)\delta}^{i\delta}\phi_j\left(\frac{t}{T}\right)\left[Z_t - Z_{(i-1)\delta}\right]dt\\ &+ \frac{1}{T}\sum_{i=1}^n \int_{(i-1)\delta}^{i\delta}\left[\phi_j\left(\frac{t}{T}\right) - \phi_j\left(\frac{i-1}{n}\right)\right]Z_{(i-1)\delta}dt + O_p\left(\frac{\delta}{T}\sup_{t\in[0,T]}\|Z_t\|\right).\end{aligned}$$

Using

$$\left\|Z_t - Z_{(i-1)\delta}\right\| \leq \left\|Z_t^c - Z_{(i-1)\delta}^c\right\| + \sum_{(i-1)\delta<\tau\leq t} \|\Delta Z_\tau\|$$

and Assumptions 3.2.1 and 3.2.2, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{i=1}^{n}\int_{(i-1)\delta}^{i\delta}\left\|\phi_j\left(\frac{t}{T}\right)\left[Z_t - Z_{(i-1)\delta}\right]\right\|dt \\
&\leq \frac{1}{T}\sum_{i=1}^{n}\int_{(i-1)\delta}^{i\delta}\left|\phi_j\left(\frac{t}{T}\right)\right|\sup_{\|\tilde{\tau}-\tau\|\leq\delta}\|Z_{\tilde{\tau}}^c - Z_\tau^c\|\,dt + \frac{1}{T}\sum_{i=1}^{n}\int_{(i-1)\delta}^{i\delta}\left|\phi_j\left(\frac{t}{T}\right)\right|\sum_{(i-1)\delta<\tau\leq i\delta}\|\Delta Z_\tau\|\,dt \\
&\leq \frac{1}{T}\sum_{i=1}^{n}\int_{(i-1)\delta}^{i\delta}\sup_{\|\tilde{\tau}-\tau\|\leq\delta}\|Z_{\tilde{\tau}}^c - Z_\tau^c\|\,dt + \frac{\delta}{T}\sum_{\tau=0}^{T}\|\Delta Z_\tau\|\max_{r\in[0,1]}\left|\phi_j(r)\right| \\
&= O_p\left(\Delta_{\delta,T}(Z)\right) + O_p(\delta).
\end{aligned}$$

In addition, for some $i^* \in (i-1, i]$,

$$\begin{aligned}
&\frac{1}{T}\sum_{i=1}^{n}\int_{(i-1)\delta}^{i\delta}\left\|\left[\phi_j\left(\frac{t}{T}\right) - \phi_j\left(\frac{i-1}{n}\right)\right]Z_{(i-1)\delta}\right\|dt \\
&\leq \frac{1}{T}\sum_{i=1}^{n}\int_{(i-1)\delta}^{i\delta}\frac{1}{n}\left|\dot{\phi}_j\left(\frac{t^*}{n}\right)\right|\left\|Z_{(i-1)\delta}\right\|dt \\
&\leq \max_{r\in[0,1]}\left|\dot{\phi}_j(r)\right|\cdot\frac{\delta}{T}\sup_{t\in[0,T]}\|Z_t\| = O_p\left(\frac{\delta}{T}\sup_{t\in[0,T]}\|Z_t\|\right),
\end{aligned}$$

where $\dot{\phi}_j(\cdot)$ is the first order derivative of $\phi_j(\cdot)$. Therefore,

$$\frac{1}{n}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)z_i - \frac{1}{T}\int_0^T\phi_j\left(\frac{t}{T}\right)Z_t dt = O_p\left(\Delta_{\delta,T}(Z) + \frac{\delta}{T}\sup_{t\in[0,T]}\|Z_t\| + \delta\right) = O_p\left(e_{\delta,T}(Z)\right).$$

■

Proof of Lemma 3.2.2. We have shown that $\sqrt{T}(\hat{\beta}_D - \beta) = \sqrt{T}(\hat{\beta}_C - \beta) + o_p(1)$. But

$$\sqrt{T}(\hat{\beta}_C - \beta) = \left[\frac{1}{T}\int_0^T X_t X_t' dt\right]^{-1}\left[\frac{1}{\sqrt{T}}\int_0^T X_t U_t dt\right] \Rightarrow S^{-1}\Omega^{1/2}W_d(1),$$

using Assumptions 3.2.3 and 3.2.4. Hence $\sqrt{T}(\hat{\beta}_D - \beta) \Rightarrow S^{-1}\Omega^{1/2}W_d(1)$.

For the second part of the lemma, we use the first part of the lemma and Lemma 3.2.1 to obtain

$$\begin{aligned}
&\frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)x_i\hat{u}_i \\
&= \frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)x_i\left[u_i - x_i'\left(\hat{\beta}_D - \beta\right)\right] \\
&= \frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)x_iu_i + \frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)x_ix_i' \cdot O_p\left(\frac{1}{\sqrt{T}}\right). \\
&= \frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)x_iu_i + \frac{1}{n}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)x_ix_i' \cdot O_p(1) \\
&= \frac{1}{\sqrt{T}}\int_0^T \phi_j\left(\frac{t}{T}\right)X_tU_t dt + o_p(1)
\end{aligned}$$

where we have used $\Lambda(n,\delta)\sqrt{T} = n$, Assumption 3.2.3, and Assumption 3.2.5(i). Under Assumption 3.2.4, we then have

$$\frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)x_i\hat{u}_i \Rightarrow \Omega^{1/2}\int_0^1 \phi_j(r)\,dW_d(r)$$

for each $j = 1,2,\ldots,K$. The joint convergence over $j = 1,2,\ldots,K$ holds by the Cramér–Wold theorem. ■

Proof of Lemma 3.3.1. Part (a). We first consider $n^{-1/2}\mathbb{W}^x\Lambda_T^{-1}$. Let $g_n : \mathbb{D}^d[0,1] \to \mathbb{D}^d[0,1]$ be defined by

$$g_n(f)(t) = \sum_{i=1}^{n} f\left(\frac{i}{n}\right)1\left\{t \in \left[\frac{i-1}{n}, \frac{i}{n}\right)\right\} + f(1)1\{t=1\}.$$

If the functions $f_n \in \mathbb{D}^d[0,1]$ are such that $f_n \to f$ for a continuous function f, then the continuity of ϕ_j in Assumption 3.2.2 implies that $\phi_j(\cdot)f_n(\cdot) \to \phi_j(\cdot)f(\cdot)$ in $\mathbb{D}^d[0,1]$ and $\phi_j(\cdot)f(\cdot)$ is a continuous function. It follows from the basic properties of the Skorokhod topology that $g_n(\phi_j f_n) \to \phi_j f$. Using the weak convergence $\Lambda_T^{-1}X_{Tr} \Rightarrow X^\circ(r)$ in Assumption 3.3.2 and the extended continuous mapping theorem (c.f. Theorem 1.11.1 of van der Vaart and Wellner (1996)), we have $g_n(\phi_j(t)\left(\Lambda_T^{-1}X_{Tt}\right)) \Rightarrow \phi_j(t)X^{\circ}(t),\ t \in [0,1]$. Combining this with the continuous mapping theorem, we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\Lambda_T^{-1}\mathbb{W}_j^x &= \frac{1}{n}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)\Lambda_T^{-1}x_i = \frac{1}{n}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)\Lambda_T^{-1}X_{i\delta}\\ &= \frac{1}{n}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)\Lambda_T^{-1}X_{\frac{i}{n}T} = \int_0^1 g_n\left(\phi_j(t)\Lambda_T^{-1}X_{Tt}\right)dt\\ &\Rightarrow \int_0^1 \phi_j(r)X^{\circ}(r)dr := \eta_j.\end{aligned}$$

This holds jointly for $j = 1,\ldots,K$ and therefore,

$$\frac{1}{\sqrt{n}}\mathbb{W}^x\Lambda_T^{-1} \Rightarrow \eta. \tag{3.18}$$

Next, under Assumption 3.3.1, Lemma 3.2.1 holds with $Z_t = U_{0t}$. Hence,

$$\begin{aligned}\sqrt{\delta}\mathbb{W}_j^{u_0} &= \frac{\sqrt{\delta}}{\sqrt{n}}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)u_{0i} = \frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)u_{0i}\\ &= \frac{1}{\sqrt{T}}\int_0^T\phi_j\left(\frac{t}{T}\right)U_{0t}dt + o_p(1).\end{aligned}$$

Let $S_t = T^{-1/2}\int_0^t U_{0r}dr$ for $t \in (0,T]$ and $S_0 = 0$. Using the continuous mapping theorem and integration by parts, we obtain, jointly for $j = 1,\ldots,K$,

$$\begin{aligned}&\sqrt{\delta}\mathbb{W}_j^{u_0} = \int_0^T\phi_j\left(\frac{t}{T}\right)dS_t + o_p(1)\\ &= \int_0^1\phi_j(r)dS_{Tr} + o_p(1) = \phi_j(1)S_T - \phi_j(0)S_0 - \int_0^1 S_{Tr}\dot{\phi}_j(r)dr + o_p(1)\\ &\Rightarrow \sigma_0\phi_j(1)W_0(1) - \sigma_0\phi_j(0)W_0(0) - \sigma_0\int_0^1\dot{\phi}_j(r)W_0(r)dr\\ &= \sigma_0\int_0^1\phi_j(r)dW_0(r),\end{aligned}$$

where the weak convergence follows from Assumption 3.3.2. Therefore,

$$\sqrt{\delta}\mathbb{W}^{u_0} \Rightarrow \nu. \tag{3.19}$$

The joint convergence of $\Lambda_T^{-1}X_{Tt}$ and $T^{-1/2}\int_0^{Tt} U_{0r}dr$ in Assumption 3.3.2 yields that (3.18) and (3.19) hold jointly, i.e., $(n^{-1/2}\mathbb{W}^x\Lambda_T^{-1}, \sqrt{\delta}\mathbb{W}^{u_0}) \Rightarrow (\eta, \nu)$.

Part (b). We write

$$\mathbb{W}^y = \mathbb{W}^x\beta_0 + \mathbb{W}^{u_0} + \alpha_0\mathbb{W}^\alpha \tag{3.20}$$

where

$$\mathbb{W}^\alpha = (\mathbb{W}_1^\alpha, \ldots, \mathbb{W}_K^\alpha)' \text{ with } \mathbb{W}_j^\alpha = \frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_j\left(\frac{i}{n}\right).$$

Note that for each $j = 1, \ldots, K$ we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^n \phi_j\left(\frac{i}{n}\right) &= \sqrt{n}\frac{1}{n}\sum_{i=1}^n \phi_j\left(\frac{i}{n}\right) \\ &= \sqrt{n}\left(\int_0^1 \phi_j(r)\,dr + O\left(\frac{1}{n}\right)\right) = O\left(\frac{1}{\sqrt{n}}\right) = o(1).\end{aligned}$$

Therefore,

$$\mathbb{W}^y = \mathbb{W}^x\beta_0 + \mathbb{W}^{u_0} + o_p(1). \tag{3.21}$$

It then follows that

$$\hat{\beta}_{TOLS} = \left(\mathbb{W}^{x\prime}\mathbb{W}^x\right)^{-1}\left(\mathbb{W}^{x\prime}\left[\mathbb{W}^x\beta_0 + \mathbb{W}^{u_0} + o_p(1)\right]\right), \tag{3.22}$$

and so

$$\hat{\beta}_{TOLS} - \beta_0 = \left(\mathbb{W}^{x\prime}\mathbb{W}^x\right)^{-1}\left(\mathbb{W}^{x\prime}\left[\mathbb{W}^{u_0} + o_p(1)\right]\right).$$

By Part (a) and Assumption 3.3.3, we then have

$$
\begin{aligned}
&\sqrt{T}\Lambda_T\left[\hat{\beta}_{TOLS}-\beta_0\right]\\
&=n^{1/2}\Lambda_T\sqrt{\delta}\left[\hat{\beta}_{TOLS}-\beta_0\right]\\
&=\left[\left(n^{1/2}\Lambda_T\right)^{-1}\left(\mathbb{W}^{x\prime}\mathbb{W}^{x}\right)\left(n^{1/2}\Lambda_T\right)^{-1}\right]^{-1}\left(n^{1/2}\Lambda_T\right)^{-1}\mathbb{W}^{x\prime}\mathbb{W}^{u_0}\sqrt{\delta}\left(1+o_p\left(1\right)\right)\\
&\Rightarrow\left(\eta'\eta\right)^{-1}\left(\eta'\nu\right).
\end{aligned}
$$

■

Proof of Theorem 3.3.1. By definition, $\hat{\mathbb{W}}^u=\mathbb{W}^y-\mathbb{W}^x\hat{\beta}_{TOLS}$. Using (3.21) and (3.22), we then have

$$
\begin{aligned}
\hat{\mathbb{W}}^u&=\mathbb{W}^x\beta_0+\mathbb{W}^{u_0}+o_p\left(1\right)-\mathbb{W}^x\left(\mathbb{W}^{x\prime}\mathbb{W}^{x}\right)^{-1}\mathbb{W}^{x\prime}\left[\mathbb{W}^x\beta_0+\mathbb{W}^{u_0}+o_p\left(1\right)\right]\\
&=\left[\mathbb{I}_K-\mathbb{W}^x\left(\mathbb{W}^{x\prime}\mathbb{W}^{x}\right)^{-1}\mathbb{W}^{x\prime}\right]\left(\mathbb{W}^{u_0}+o_p\left(1\right)\right). \qquad (3.23)
\end{aligned}
$$

Hence, by Lemma 3.3.1(i),

$$
\begin{aligned}
\delta\cdot\hat{\sigma}_0^2&=\frac{1}{K}\sqrt{\delta}\left(\mathbb{W}^{u_0}+o_p\left(1\right)\right)'\left[\mathbb{I}_K-\mathbb{W}^x\left(\mathbb{W}^{x\prime}\mathbb{W}^{x}\right)^{-1}\mathbb{W}^{x\prime}\right]\sqrt{\delta}\left(\mathbb{W}^{u_0}+o_p\left(1\right)\right)\\
&\Rightarrow\frac{1}{K}\nu'M_\eta\nu.
\end{aligned}
$$

where $M_\eta=\mathbb{I}_K-\eta(\eta'\eta)^{-1}\eta'$. Using Lemma 3.3.1(ii), we have, under H_0,

$$
\sqrt{T}\tilde{\Lambda}_T(R\hat{\beta}_{TOLS}-r)=(\tilde{\Lambda}_TR\Lambda_T^{-1})\sqrt{T}\Lambda_T(\hat{\beta}_{TOLS}-\beta_0)\Rightarrow R_\circ\left(\eta'\eta\right)^{-1}\left(\eta'\nu\right)
$$

and

$$\begin{aligned} & n\tilde{\Lambda}_T^{-1}\left[R\left(\mathbb{W}^{x\prime}\mathbb{W}^x\right)^{-1}R'\right]^{-1}\tilde{\Lambda}_T^{-1} \\ &= n\tilde{\Lambda}_T^{-1}\left\{Rn^{1/2}\Lambda_T^{-1}\left[\left(\mathbb{W}^x\Lambda_T^{-1}n^{-1/2}\right)'\mathbb{W}^x\Lambda_T^{-1}n^{-1/2}\right]^{-1}\left(Rn^{1/2}\Lambda_T^{-1}\right)'\right\}^{-1}\tilde{\Lambda}_T^{-1} \\ &= \left\{\tilde{\Lambda}_T R\Lambda_T^{-1}\left[\left(\mathbb{W}^x\Lambda_T^{-1}n^{-1/2}\right)'\mathbb{W}^x\Lambda_T^{-1}n^{-1/2}\right]^{-1}\left(\tilde{\Lambda}_T R\Lambda_T^{-1}\right)'\right\}^{-1} \\ &\Rightarrow [R_\circ\left(\eta'\eta\right)^{-1}R_\circ']^{-1}. \end{aligned}$$

Therefore,

$$\begin{aligned} F_{TOLS} &= \frac{1}{\hat{\sigma}_0^2}(R\hat{\beta}_{TOLS}-r)'\left[R\left(\mathbb{W}^{x\prime}\mathbb{W}^x\right)^{-1}R'\right]^{-1}(R\hat{\beta}_{TOLS}-r)/p \\ &= \frac{1}{p}\frac{1}{\delta\hat{\sigma}_0^2}(R\hat{\beta}_{TOLS}-r)'\sqrt{T}\tilde{\Lambda}_T \\ &\times n\tilde{\Lambda}_T^{-1}\left[R\left(\mathbb{W}^{x\prime}\mathbb{W}^x\right)^{-1}R'\right]^{-1}\tilde{\Lambda}_T^{-1}\times\sqrt{T}\tilde{\Lambda}_T(R\hat{\beta}_{TOLS}-r) \\ &\Rightarrow \frac{K}{p}\frac{\left[R_\circ(\eta'\eta)^{-1}\eta'\nu\right]'\left(R_\circ\left(\eta'\eta\right)^{-1}R_\circ'\right)^{-1}\left[R_\circ(\eta'\eta)^{-1}\eta'\nu\right]}{\nu'M_\eta\nu} \\ &= \frac{K}{p}\frac{Q'\left(R_\circ\left(\eta'\eta\right)^{-1}R_\circ'\right)^{-1}Q}{\nu'M_\eta\nu/\sigma_0^2}, \end{aligned} \tag{3.24}$$

where $Q = R_\circ(\eta'\eta)^{-1}\eta'\nu/\sigma_0$. Now, conditional on η,

$$Q'\left(R_\circ\left(\eta'\eta\right)^{-1}R_\circ'\right)^{-1}Q \overset{d}{=} \chi_p^2, \text{ and } \nu'M_\eta\nu/\sigma_0^2 \overset{d}{=} \chi_{K-d}^2.$$

Additionally, conditional on η, $M_\eta\nu$ and $\eta'\nu$ are independent, as both $M_\eta\nu$ and $\eta'\nu$ are normal and the conditional covariance is

$$cov\left(M_\eta\nu,\eta'\nu\right) = M_\eta\eta = 0.$$

Thus, conditional on η, the numerator and the denominator in (3.24) are independent chi-squared variates. This implies that

$$\frac{K}{p}\frac{Q'\left(R_\circ\left(\eta'\eta\right)^{-1}R_\circ'\right)^{-1}Q}{\nu' M_\eta \nu/\sigma_0^2} = \frac{K}{K-d}\frac{Q'\left(R_\circ\left(\eta'\eta\right)^{-1}R_\circ'\right)^{-1}Q/p}{\nu' M_\eta \nu/\left[\sigma_0^2(K-d)\right]} \overset{d}{=} \frac{K}{K-d}F_{p,K-d}$$

conditional on η. But the conditional distribution does not depend on the conditioning variable η, so it is also the unconditional distribution. This proves the second statement of the theorem. ■

Proof of Theorem 3.3.2. Part (a): Setting $\Lambda_T = \sqrt{T}\mathbb{I}_d$ and $X^\circ(r) = B_x(r)$ we can proceed nearly identically to the proof of Lemma 3.3.1(a) to obtain that

$$\left[(nT)^{-1/2}\mathbb{W}^x, \delta^{1/2}\mathbb{W}^{u_0}\right] \Rightarrow (\eta, \nu).$$

It remains to show that $\delta^{1/2}\mathbb{W}^{\tilde{\Delta}x} \Rightarrow \xi$ jointly with the above convergence. The joint convergence holds by the Cramér–Wold theorem. It remains to prove the marginal convergence $\delta^{1/2}\mathbb{W}^{\tilde{\Delta}x} \Rightarrow \xi$. We have

$$\begin{aligned}
\delta^{1/2}\mathbb{W}_j^{\tilde{\Delta}x} &= \frac{1}{\sqrt{n\delta}}\sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)[x_i - x_{i-1}] = \sum_{i=1}^{n}\phi_j\left(\frac{i}{n}\right)T^{-1/2}[x_i - x_{i-1}] \\
&= \frac{1}{n}\sum_{i=1}^{n-1}\frac{\left[\phi_j\left(\frac{i}{n}\right) - \phi_j\left(\frac{i+1}{n}\right)\right]}{1/n}T^{-1/2}x_i + \phi_j(1)T^{-1/2}x_n - \phi_j\left(\frac{1}{n}\right)T^{-1/2}x_0 \\
&= -\frac{1}{n}\sum_{i=1}^{n-1}\dot{\phi}_j\left(\frac{i}{n}\right)T^{-1/2}X_{\frac{i}{n}T} + \phi_j(1)T^{-1/2}X_T - \phi_j\left(\frac{1}{n}\right)T^{-1/2}X_0 \\
&\quad + O_p\left(\frac{1}{n}\frac{1}{n}\sum_{i=1}^{n-1}T^{-1/2}\left\|X_{\frac{i}{n}T}\right\|\right). \qquad (3.25)
\end{aligned}$$

Using the continuous mapping theorem and Assumption 3.3.4, we have

$$n^{-1}\sum_{i=1}^{n-1}T^{-1/2}\left\|X_{\frac{i}{n}T}\right\| \Rightarrow \int_0^1 \|B_x(r)\|\,dr$$

and hence the last term in (3.25) is of order $O_p(1/n) = o_p(1)$. Therefore, using integration by parts,

$$\begin{aligned}
&\delta^{1/2}\mathbb{W}_j^{\tilde{\Delta}x} \\
&= \frac{1}{n}\sum_{i=1}^{n-1}\dot{\phi}_j\left(\frac{i}{n}\right)T^{-1/2}X_{\frac{i}{n}T} + \phi_j(1)T^{-1/2}X_T - \phi_j\left(\frac{1}{n}\right)T^{-1/2}X_0 + o_p(1) \\
&\Rightarrow -\int_0^1 \dot{\phi}_j(r)B_x(r)dr + \phi_j(1)B_x(1) - \phi_j(0)B_x(0) \\
&= \int_0^1 \phi_j(r)dB_x(r) = \xi_j.
\end{aligned}$$

This holds jointly for $j = 1, \ldots, K$ so that $\delta^{1/2}\mathbb{W}^{\tilde{\Delta}x} \Rightarrow \xi$.

Part (b). Following the same argument as in the proof Theorem 3.3.1, we can ignore the intercept. To simplify the notation, we assume from the outset that there is no intercept in the model so that

$$\mathbb{W}^y = \mathbb{W}^x\beta_0 + \mathbb{W}^{u_0}.$$

Given this, we have

$$\hat{\gamma} - \begin{pmatrix} \beta_0 \\ 0 \end{pmatrix} = \begin{pmatrix} \mathbb{W}^{x\prime}\mathbb{W}^x & \mathbb{W}^{x\prime}\mathbb{W}^{\tilde{\Delta}x} \\ \mathbb{W}^{\tilde{\Delta}x\prime}\mathbb{W}^x & \mathbb{W}^{\tilde{\Delta}x\prime}\mathbb{W}^{\tilde{\Delta}x} \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{W}^{x\prime}\mathbb{W}^{u_0} \\ \mathbb{W}^{\tilde{\Delta}x\prime}\mathbb{W}^{u_0} \end{pmatrix}.$$

Recall that $\Upsilon_T = diag\left(T\mathbb{I}_d, \mathbb{I}_d\right)$. Using Part (a) and noting that $\delta^{1/2}/T = (nT)^{-1/2}$, we have

$$\begin{aligned}
&\Upsilon_T \left[\hat{\gamma} - \begin{pmatrix} \beta_0 \\ 0 \end{pmatrix}\right] \\
&= \left[\delta^{1/2}\Upsilon_T^{-1} \begin{pmatrix} (\mathbb{W}^x)'\,\mathbb{W}^x & \mathbb{W}^{x\prime}\mathbb{W}^{\tilde{\Delta}x} \\ (\mathbb{W}^{\tilde{\Delta}x})'\mathbb{W}^x & \mathbb{W}^{\tilde{\Delta}x\prime}\mathbb{W}^{\tilde{\Delta}x} \end{pmatrix} \Upsilon_T^{-1}\delta^{1/2}\right]^{-1} \Upsilon_T^{-1}\delta^{1/2} \begin{pmatrix} \mathbb{W}^{x\prime}\mathbb{W}^{u_0}\delta^{1/2} \\ \mathbb{W}^{\tilde{\Delta}x\prime}\mathbb{W}^{u_0}\delta^{1/2} \end{pmatrix} \\
&= \begin{pmatrix} (nT)^{-1/2}\,\mathbb{W}^{x\prime}\mathbb{W}^x\,(nT)^{-1/2} & (nT)^{-1/2}\,\mathbb{W}^{x\prime}\mathbb{W}^{\tilde{\Delta}x}\delta^{1/2} \\ \delta^{1/2}\mathbb{W}^{\tilde{\Delta}x}\mathbb{W}^x\,(nT)^{-1/2} & \delta^{1/2}\mathbb{W}^{\tilde{\Delta}x\prime}\mathbb{W}^{\tilde{\Delta}x}\delta^{1/2} \end{pmatrix}^{-1} \begin{pmatrix} (nT)^{-1/2}\,\mathbb{W}^{x\prime}\mathbb{W}^{u_0}\delta^{1/2} \\ \delta^{1/2}\mathbb{W}^{\tilde{\Delta}x\prime}\mathbb{W}^{u_0}\delta^{1/2} \end{pmatrix} \\
&\Rightarrow \begin{pmatrix} \eta'\eta & \eta'\xi \\ \xi'\eta & \xi'\xi \end{pmatrix}^{-1} \begin{pmatrix} \eta'\nu \\ \xi'\nu \end{pmatrix}.
\end{aligned}$$

Plugging $\nu = \sigma_{0\cdot x}\tilde{\nu} + \xi\,\theta_0$ into the above limit, we have

$$\begin{aligned}
\Upsilon_T \left[\hat{\gamma} - \begin{pmatrix} \beta_0 \\ 0 \end{pmatrix}\right] &\Rightarrow \begin{pmatrix} \eta'\eta & \eta'\xi \\ \xi'\eta & \xi'\xi \end{pmatrix}^{-1} \begin{pmatrix} \eta'\xi \\ \xi'\xi \end{pmatrix} \theta_0 + \sigma_{0\cdot x} \begin{pmatrix} \eta'\eta & \eta'\xi \\ \xi'\eta & \xi'\xi \end{pmatrix}^{-1} \begin{pmatrix} \eta'\tilde{\nu} \\ \xi'\tilde{\nu} \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ \theta_0 \end{pmatrix} + \sigma_{0\cdot x}\left(\zeta'\zeta\right)^{-1}\zeta'\tilde{\nu}.
\end{aligned}$$

That is, $\Upsilon_T\left(\hat{\gamma} - \gamma_0\right) \Rightarrow \sigma_{0\cdot x}\left(\zeta'\zeta\right)^{-1}\zeta'\tilde{\nu}$. The first block of this result is $T(\hat{\beta}_{TAOLS} - \beta_0) \Rightarrow \sigma_{0\cdot x}\left(\eta' M_\xi \eta\right)^{-1}\eta' M_\xi \tilde{\nu}$.

Part (c). First, it follows from Part (b) that under H_0,

$$T(R\hat{\beta}_{TAOLS} - r) = RT(\hat{\beta}_{TAOLS} - \beta_0) \Rightarrow \sigma_{0\cdot x}R\left(\eta' M_\xi \eta\right)^{-1}\eta' M_\xi \tilde{\nu}. \tag{3.26}$$

Next,

$$
\begin{aligned}
&\delta(\hat{\mathbb{W}}^{u_{0\cdot x}})'\hat{\mathbb{W}}^{u_{0\cdot x}} \\
&= \delta \mathbb{W}^{u_0\prime}\left[I-\widetilde{\mathbb{W}}(\widetilde{\mathbb{W}}'\widetilde{\mathbb{W}})^{-1}\widetilde{\mathbb{W}}'\right]\mathbb{W}^{u_0} \\
&= (\sqrt{\delta}\mathbb{W}^{u_0})'\left[I-(\widetilde{\mathbb{W}}\delta^{1/2}\Upsilon_T^{-1})\left[(\widetilde{\mathbb{W}}\delta^{1/2}\Upsilon_T^{-1})(\widetilde{\mathbb{W}}\delta^{1/2}\Upsilon_T^{-1})'\right]^{-1}(\widetilde{\mathbb{W}}\delta^{1/2}\Upsilon_T^{-1})'\right]\sqrt{\delta}\mathbb{W}^{u_0} \\
&\Rightarrow \nu'\left(I-\zeta\left(\zeta'\zeta\right)^{-1}\zeta'\right)\nu = \sigma_{0\cdot x}^2\tilde{\nu}'\left(I-\zeta\left(\zeta'\zeta\right)^{-1}\zeta'\right)\tilde{\nu} = \sigma_{0\cdot x}^2\tilde{\nu}'M_\zeta\tilde{\nu}.
\end{aligned}
$$

Hence,

$$
\delta\hat{\sigma}_{0\cdot x}^2 = \frac{1}{K}\delta(\hat{\mathbb{W}}^{u_{0\cdot x}})'\hat{\mathbb{W}}^{u_{0\cdot x}} \Rightarrow \frac{1}{K}\sigma_{0\cdot x}^2\tilde{\nu}'M_\zeta\tilde{\nu}. \tag{3.27}
$$

Combining (3.26) and (3.27), we have

$$
\begin{aligned}
&F_{TAOLS} \\
&= \frac{1}{\hat{\sigma}_{0\cdot x}^2}(R\hat{\beta}_{TAOLS}-r)'\left[R\left(\mathbb{W}^{x\prime}M_{\tilde{\Delta}x}\mathbb{W}^x\right)^{-1}R'\right]^{-1}(R\hat{\beta}_{TAOLS}-r)/p \\
&= \frac{1}{p\delta\hat{\sigma}_{0\cdot x}^2}[RT(\hat{\beta}_{TAOLS}-\beta_0)]'\{R[(nT)^{-1/2}\mathbb{W}^{x\prime}M_{\tilde{\Delta}x}\mathbb{W}^x(nT)^{-1/2}]^{-1}R'\}^{-1}RT(\hat{\beta}_{TAOLS}-\beta_0) \\
&\Rightarrow \frac{\left[R\left(\eta'M_\xi\eta\right)^{-1}\eta'M_\xi\tilde{\nu}\right]'\left[R\left(\eta'M_\xi\eta\right)^{-1}R'\right]^{-1}\left[R\left(\eta'M_\xi\eta\right)^{-1}\eta'M_\xi\tilde{\nu}\right]/p}{\tilde{\nu}'M_\zeta\tilde{\nu}/K} \\
&= \frac{K}{p}\frac{Q'\left(R\left(\eta'M_\xi\eta\right)^{-1}R'\right)^{-1}Q}{\tilde{\nu}'M_\zeta\tilde{\nu}},
\end{aligned}
$$

where $Q = R(\eta'M_\xi\eta)^{-1}\eta'M_\xi\tilde{\nu}$. Following the argument similar to that in the proof of Theorem 3.3.1, we can then show that $F_{TAOLS} \Rightarrow \frac{K}{K-2d}\cdot F_{p,K-2d}$. ■

3.B Supplementary Appendix for Chapter 3

In this appendix, we develop an MSE-optimal rule for choosing K. Part of our theoretical analysis is the high-frequency continuous-time counterparts of Phillips (2005), which develops a rule for choosing K in LRV estimation for a fully observed discrete-time process. We allow for more general basis functions while Phillips (2005) considers only sine and cosine basis functions. Thus, even for usual discrete-time processes, our theoretical development goes beyond Phillips (2005).

3.B.1 MSE-optimal Choice of K

To abstract away the technical issues that will not affect the practical implementation of the proposed rule, we define the infeasible variance estimator as in the main text:

$$\hat{\Omega}^{*}=\frac{1}{K}\sum_{j=1}^{K}\left[\frac{1}{\Lambda(n,\delta)}\sum_{i=1}^{n}\phi_{j}\left(\frac{i}{n}\right)(x_{i}u_{i})\right]^{\otimes 2}.$$

$\hat{\Omega}^{*}$ is infeasible because u_i is not observed. We choose K to minimize the asymptotic MSE of $\hat{\Omega}^{*}$. We could alternatively follow Andrews (1991) to find the approximate and truncated MSE of the feasible estimator $\hat{\Omega}$ and use it to guide the choice of K. These two approaches will lead to the same formula for the MSE-optimal K. Here we opt for the simpler approach.

Assumption 3.B.1 *The following hold:*

(i)

$$var\left[\mathrm{vec}(\hat{\Omega}^{*})\right]=var\left[\mathrm{vec}\left(\Omega^{1/2}\frac{1}{K}\sum_{j=1}^{K}\left[\int_{0}^{1}\phi_{j}(r)\,dW_{d}(r)\right]^{\otimes 2}\Omega^{1/2}\right)\right](1+o(1))$$

as $T\rightarrow\infty$ for both a fixed K and a growing K $(i.e.,$ $K\rightarrow\infty)$.

(ii) *Let* $\Gamma_{XU}(\tau)=E\left(X_{t}U_{t}U_{t-\tau}X_{t-\tau}'\right).$ *For some* $\iota>0,$ *there exists positive constants* C_1

and C_2 *such that*

$$\|\Gamma_{XU}(\tau)\| \leq C_1 \textit{ for all } \tau \textit{ and } \|\Gamma_{XU}(\tau)\| \leq C_1 \tau^{-(3+\iota)} \textit{ for all } |\tau| \geq C_2.$$

(iii) $\delta \sum_{k=-n+1}^{n-1} (k\delta)^m \Gamma_{XU}(k\delta) - \int_{-T}^{T} \tau^m \Gamma_{XU}(\tau)\, d\tau = O(\delta)$ *for* $m = 0, 2.$

(iv) *For some constant* $C > 0$, $\sup_{j\in[K]} \sup_{r\in[0,1]} \max\left\{\left|\phi_j(r)\right|, \left|\dot{\phi}_j(r)\right|/j\right\} \leq C$ *where* $\dot{\phi}_j$ *is the first order derivative of* ϕ_j *and* $[K] := \{1, \ldots, K\}$.

(v) *If* $K \to \infty$ *as* $T \to \infty$, *then, for some constant* $c_{\phi,2} \neq 0$,

$$\lim_{K\to\infty}\left[-\frac{1}{K^3}\sum_{j=1}^{K}\frac{1}{2}\int_0^1 \phi_j(r)\ddot{\phi}_j(r)\,dr\right] = c_{\phi,2},$$

where $\ddot{\phi}_j$ *is the second order derivative of* ϕ_j.

Assumption 3.B.1(i) is a high-level assumption. When K is fixed and Assumptions 3.2.1–3.2.5 hold,

$$\hat{\Omega}^* \Rightarrow \Omega^{1/2}\frac{1}{K}\sum_{j=1}^{K}\left[\int_0^1 \phi_j(r)\,dW_d(r)\right]^{\otimes 2}\Omega^{1/2}.$$

So Assumption 3.B.1(i) says that the limit of the exact finite sample variance of vec$(\hat{\Omega}^*)$ is equal to the variance of its limiting distribution, namely the asymptotic variance. From a theoretical point of view, this is plausible if we have enough moment conditions. Alternatively, we simply use the asymptotic variance in place of the exact finite sample variance to obtain an approximate MSE. This is, in fact, a typical approach for smoothing parameter choice in a nonparametric setting when the exact finite sample variance is difficult, if not impossible, to obtain. For both a fixed K and a growing K, we can show that an assumption similar to Assumption 2.3(b) in Lu and Park (2019), Assumption 3.2.2, and Assumptions 3.B.1(ii)-(iv) are sufficient for Assumption 3.B.1(i). The details and proof are given in the supplementary appendix.

Assumption 3.B.1(ii) imposes that the covariance $\|\Gamma_{XU}(\tau)\|$ is bounded above and decays to zero at a certain rate. The assumption ensures that $\delta\sum_{k=-\infty}^{\infty}(k\delta)^2\|\Gamma_{XU}(k\delta)\| < \infty$

and $\int_{-\infty}^{\infty} \tau^2 \|\Gamma_{XU}(\tau)\| < \infty$ (see the proof Theorem 3.B.1). The summability condition can be regarded as the continuous counterpart of the integrability condition. These conditions are often imposed directly in the literature. For the latter condition, see, for example, Assumption 2.2 in Lu and Park (2019) (pp. 239).

Assumption 3.B.1(iii) assumes that the discrete sum is a good approximation to the integral. Note that

$$\begin{aligned}
&\delta \sum_{k=-n+1}^{n-1} (k\delta)^m \Gamma_{XU}(k\delta) - \int_{-T}^{T} \tau^m \Gamma_{XU}(\tau)\, d\tau \\
&= \sum_{k=-n+1}^{n-1} \left[\int_{k\delta}^{(k+1)\delta} \left[(k\delta)^m \Gamma_{XU}(k\delta) - \tau^m \Gamma_{XU}(\tau) \right] d\tau \right] + O(\delta) \\
&= \left[\sum_{k=-n+1}^{n-1} \max_{t\in[k\delta,(k+1)\delta]} \frac{\partial \left[t^m \Gamma_{XU}(t) \right]}{\partial t} \delta + O(1) \right] \delta.
\end{aligned}$$

Therefore, Assumption 3.B.1(iii) holds if $\delta \sum_{k=-n+1}^{n-1} \max_{t\in[k\delta,(k+1)\delta]} \left\| \frac{\partial [t^m \Gamma_{XU}(t)]}{\partial t} \right\| < \infty$.

Assumptions 3.B.1(iv) and (v) contain some additional mild conditions on the basis functions. The assumptions are satisfied for the sine and cosine basis functions (i.e., Fourier bases) given in (3.3). For this set of Fourier bases, we have

$$\ddot{\phi}_{2j-1}(r) = -\sqrt{2}(2\pi j)^2 \cos(2\pi jr) \text{ and } \ddot{\phi}_{2j}(r) = -\sqrt{2}(2\pi j)^2 \sin(2\pi jr) \text{ for } j = 1, \ldots, K/2,$$

and hence

$$\begin{aligned}
c_{\phi,2} &= -\lim_{K\to\infty} \frac{1}{K^3} \sum_{j=1}^{K} \frac{1}{2} \int_0^1 \phi_j(r)\, \ddot{\phi}_j(r)\, dr \\
&= \lim_{K\to\infty} \frac{1}{K^3} \sum_{j=1}^{K/2} \frac{4\pi^2 j^2}{2} \left[\int_0^1 2\sin(2\pi jr)^2\, dr + \int_0^1 2\cos(2\pi jr)^2\, dr \right] \\
&= \lim_{K\to\infty} \frac{1}{K^3} \sum_{j=1}^{K/2} 4\pi^2 j^2 = \int_0^{1/2} 4\pi^2 x^2 dx = \frac{\pi^2}{6}.
\end{aligned}$$

For a kernel function $k(\cdot)$ with Parzen exponent q, the asymptotic bias of the kernel LRV estimator depends on the "Parzen parameter" $c_{k,q}$ defined by

$$c_{k,q} = \lim_{x \to 0} \frac{1 - k(x)}{x^q}.$$

The parameter $c_{\phi,2}$ in Assumption 3.B.1(v) plays the same role in series LRV estimation as $c_{k,q}$ does in kernel LRV estimation. Here, the assumptions imposed on the basis functions ensure that the resulting series LRV estimator is analogous to a kernel LRV estimator with a second-order kernel (i.e., its Parzen exponent q is equal to 2). There are other sets of basis functions such as Legendre polynomials that deliver series LRV estimators with asymptotic properties similar to the kernel LRV estimators based on a first-order kernel (e.g., the Bartlett kernel). See Lazarus et al. (2018) for more discussion. Hwang and Sun (2018) discusses why the set of Legendre polynomials may not be a good choice. We focus on second-order series LRV estimators in this paper.

Theorem 3.B.1 *Let Assumption 3.B.1 hold.*

(a) Under Assumption 3.B.1(i), as $T \to \infty$, the variance of $\hat{\Omega}^*$ satisfies

$$\operatorname{var}\left[\operatorname{vec}(\hat{\Omega}^*)\right] = \frac{1}{K}(\Omega \otimes \Omega)(\mathbb{I}_{d^2} + \mathbb{K}_{dd})(1 + o(1)),$$

where $\mathbb{I}_{d^2}$ is the $d^2 \times d^2$ identity matrix and $\mathbb{K}_{dd}$ is the $d^2 \times d^2$ commutation matrix.

(b) Under Assumptions 3.B.1(ii)–(v), as $T \to \infty$ and $K \to \infty$, the bias of $\hat{\Omega}^*$ satisfies

$$E(\hat{\Omega}^* - \Omega) = -c_{\phi,2}\frac{K^2}{T^2}B_2 + o\left(\frac{K^2}{T^2}\right) + O\left(\delta + \frac{(\log n)^2}{T^2} + \frac{1}{T}\right),$$

where

$$B_2 = \int_{-\infty}^{\infty} \tau^2 \Gamma_{XU}(\tau)\, d\tau.$$

(c) Under Assumptions 3.B.1(ii)–(iv), as $T \to \infty$ for a fixed K, the bias of $\hat{\Omega}^*$ satisfies

$$E(\hat{\Omega}^* - \Omega) = -\frac{1}{T} c_{\phi,1} B_1 + o\left(\frac{1}{T}\right) + O\left(\delta + \frac{(\log n)^2}{T^2} + \frac{1}{n}\right),$$

where

$$c_{\phi,1} = c_{\phi,1}(K) := \frac{1}{2}\frac{1}{K}\sum_{j=1}^{K}\left[\phi_j^2(1) + \phi_j^2(0)\right] \text{ and } B_1 = \int_{-\infty}^{\infty} \tau \Gamma_{XU}(\tau)\, d\tau.$$

When $K \to \infty$ and $T \to \infty$, the variance and bias expressions are similar to those in the case with discrete-time data. Their interpretations are also similar. For example, when $X_t U_t$ is positively autocorrelated such that $\Gamma_{XU}(\tau) > 0$ for all τ, then $B_2 > 0$ and $\hat{\Omega}^*$ is biased downward. This is analogous to the discrete-time case. Note that the dominating bias is equal to $-c_\phi K^2 T^{-2} B_2$ instead of $-c_\phi K^2 n^{-2} B_2$. The latter can be shown to be the dominating bias in the usual discrete-time case for a fixed time interval (e.g., $\delta = 1$) with n observations. A takeaway from this comparison is that the effective sample size of a high-frequency sample (i.e., $\delta \to 0$) from a continuous-time process is the time span T instead of the number of observations n over this time span. When we use the effective sample size T in the bias expression, the asymptotic bias depends only on B_2, which is an intrinsic feature of the continuous-time process. In particular, the asymptotic bias does not depend on δ. This may appear counter-intuitive. We may argue that the process becomes more persistent for a smaller δ, and so we expect a larger absolute bias for a smaller δ. Such an argument is valid if we represent the asymptotic bias in terms of n, namely $-c_\phi \left(K^2 n^{-2}\right)\left(B_2 \delta^{-2}\right)$. Smaller δ indeed leads to a larger bias for a given n, but n becomes larger for a smaller δ. The net effect is that the asymptotic bias depends on the effective sample size T but not n or δ separately.

Define[6]

$$\text{MSE}(\hat{\Omega}^*) = E\left[\text{vec}(\hat{\Omega}^* - \Omega)' \text{vec}(\hat{\Omega}^* - \Omega)\right],$$

[6] It is possible to weigh different elements of $\text{vec}(\hat{\Omega}^* - \Omega)$ differently by defining

$$\text{MSE}(\hat{\Omega}^*) = E\left[\text{vec}(\hat{\Omega}^* - \Omega)' \mathscr{W} \text{vec}(\hat{\Omega}^* - \Omega)\right]$$

for some matrix $\mathscr{W}$. Here we have implicitly chosen $\mathscr{W}$ to be an identity matrix.

which is the mean square error of vec($\hat{\Omega}^*$). It follows from Theorems 3.B.1 (i) and (ii) that

$$\begin{aligned}
&\mathrm{MSE}(\hat{\Omega}^*) \\
&= \mathrm{tr}\left[\{\Omega \otimes \Omega\}\left(\mathbb{I}_{d^2} + \mathbb{K}_{dd}\right)\right]\frac{1}{K} + c_{\phi,2}^2 \mathrm{vec}\left(B_2\right)' \mathrm{vec}\left(B_2\right)\frac{K^4}{T^4} \\
&+ o\left(\frac{1}{K} + \frac{K^4}{T^4}\right) + O\left(\delta^2 + \frac{(\log n)^4}{T^4} + \frac{1}{T^2}\right).
\end{aligned}$$

Ignoring the terms that will be shown to be of a smaller order and optimizing MSE($\hat{\Omega}^*$) over K, we obtain the formula[7]

$$K = \kappa\left(\Omega, B_2\right)^{1/5} T^{4/5}, \tag{3.28}$$

where

$$\kappa\left(\Omega, B_2\right) := \left(\frac{tr\left[\{\Omega \otimes \Omega\}\left(\mathbb{I}_{d^2} + \mathbb{K}_{dd}\right)\right]}{4c_{\phi,2}^2 \mathrm{vec}\left[B_2\right]' \mathrm{vec}\left[B_2\right]}\right).$$

When $K = \kappa\left(\Omega, B_2\right)^{1/5} T^{4/5}$, the first two terms in MSE($\hat{\Omega}^*$) are of order $T^{-4/5}$. To ensure the terms that we ignore are indeed of a smaller order, we require that

$$\delta^2 + \frac{(\log n)^4}{T^4} + \frac{1}{T^2} = o\left(T^{-4/5}\right).$$

If we set $\delta = O(T^{-\tau})$, then we require τ to be large enough. Such a requirement is compatible with the sufficient conditions for Assumption 3.2.5(i).

In the case of usual discrete time series data with a fixed sampling time interval and n observations, the optimal choice of K is given by

$$K_D = \kappa\left(\Omega_D, B_{2D}\right)^{1/5} n^{4/5}, \tag{3.29}$$

[7] Given that K is an integer, we should round $\kappa\left(\Omega, B_2\right)^{1/5} T^{4/5}$ up to the next integer and use it as K. We ignore this for the theoretical analysis but implement it in the simulation study.

where

$$\kappa(\Omega_D, B_{2D}) := \frac{tr\left[\{\Omega_D \otimes \Omega_D\}\left(\mathbb{I}_{d^2} + \mathbb{K}_{dd}\right)\right]}{4c_{\phi,2}^2 \text{vec}\left[B_{2D}\right]' \text{vec}\left[B_{2D}\right]}.$$

The formula is the same as that in (3.28) but with T replaced by n. See, for example, Phillips (2005). In the above, Ω_D and B_{2D} are the discrete analogues of Ω and B_2. If we use the formula for K in (3.29) and set $K = cn^{4/5}$ for some constant $c > 0$, then we obtain a sub-optimal rate of K for the high-frequency data with a shrinking sample interval (i.e., $\delta \to 0$). The choice of $K = cn^{4/5}$ is too large for high-frequency data. For this type of data, the neighboring observations are highly correlated, and a smaller K is desired.

Now suppose we pretend that $\{z_i = x_i u_i\}_{i=1}^n$ is a discrete-time process with a fixed time interval (e.g., $\delta = 1$) and n observations, and we use a parametric AR(1) plug-in approach to implement (3.29). As in the main text, we fit an AR(1) model to each component $z_{i,j}$ of z_i :

$$z_{i,j} = \rho_j z_{i-1,j} + e_{zj} \text{ for } j = 1, 2, \ldots, d$$

with the AR parameter and error variance estimated by

$$\hat{\rho}_j = \frac{\sum_{i=2}^n z_{i,j} z_{i-1,j}}{\sum_{i=2}^n z_{i-1,j}^2} \text{ and } \hat{\sigma}_j^2 = \frac{1}{n} \sum_{i=2}^n \left(z_{i,j} - \hat{\rho}_j z_{i-1,j}\right)^2.$$

We then compute

$$\hat{\kappa}_D = \frac{1}{8c_{\phi,2}^2} \left(\sum_{j=1}^d \frac{\hat{\rho}_j^2 \hat{\sigma}_j^4}{\left(1 - \hat{\rho}_j\right)^8} \right)^{-1} \left(\sum_{j=1}^d \frac{\hat{\sigma}_j^4}{\left(1 - \hat{\rho}_j\right)^4} \right)$$

and let

$$\hat{K}_D = \hat{\kappa}_D^{1/5} n^{4/5}. \tag{3.30}$$

The above data-driven choice does not require the value of δ, and hence we do not need to pin down the unit of time in measuring the sampling intervals. Whether the length of the sampling intervals is measured in seconds, hours, days, or months does not affect how we compute $\hat{K}_D$. The value of $\hat{K}_D$ is invariant to the unit of time, and an applied researcher does not have to choose a unit of time.

The question is whether the so-obtained $\hat{K}_D$ is of the optimal order $T^{4/5}$ with probability approaching one. On the surface, the answer is no, as $\hat{K}_D$ is apparently of order $n^{4/5}$. However, under the AR(1) plug-in implementation, $\hat{\kappa}_D$ is not a fixed constant. In fact, following Chang et al. (2021) (Lemma 4.2), we can show that as $\delta \to 0$ and $T \to \infty$,

$$\hat{\rho}_j = 1 - c_{1j}\delta + o_p(\delta) \text{ and } \hat{\sigma}_j^2 = c_{2j}\delta + o_p(\delta)$$

for some constants $c_{1j} > 0$ and $c_{2j} > 0$. Essentially, $\{z_{i,j}\}$ is a highly persistent process with the autocorrelation approaching unity at the rate of δ. The smaller δ is, the higher the autocorrelation is. As $\delta \to 0$, $\{z_{i,j}\}$ is effectively a near unit root process with the innovation variance proportional to the sampling interval δ. Plugging the above results into $\hat{\kappa}_D$ yields

$$\begin{aligned}\hat{\kappa}_D &= \frac{1}{8c_\phi^2}\left(\sum_{j=1}^d \frac{(c_{2j})^2\delta^2}{(c_{1j}\delta)^8}\right)^{-1}\left(\sum_{j=1}^d \frac{(c_{2j})^2\delta^2}{(c_{1j}\delta)^4}\right)(1+o_p(1))\\ &= \frac{1}{8c_\phi^2}\left(\sum_{j=1}^d \frac{c_{2j}^2}{c_{1j}^8}\right)^{-1}\left(\sum_{j=1}^d \frac{c_{2j}^2}{c_{1j}^4}\right)\delta^4(1+o_p(1)).\end{aligned}$$

As a result,

$$\begin{aligned}\hat{K}_D = \hat{\kappa}_D^{1/5}n^{4/5} &= \left[\frac{1}{8c_{\phi,2}^2}\left(\sum_{j=1}^d \frac{c_{2j}^2}{c_{1j}^8}\right)^{-1}\left(\sum_{j=1}^d \frac{c_{2j}^2}{c_{1j}^4}\right)\right]^{1/5}\delta^{4/5}n^{4/5}(1+o_p(1))\\ &= \left[\frac{1}{8c_{\phi,2}^2}\left(\sum_{j=1}^d \frac{c_{2j}^2}{c_{1j}^8}\right)^{-1}\left(\sum_{j=1}^d \frac{c_{2j}^2}{c_{1j}^4}\right)\right]^{1/5} T^{4/5}(1+o_p(1)).\end{aligned}$$

With probability approaching one, the rate of $\hat{K}_D$ is the same as the optimal rate of $T^{4/5}$. So the AR(1) plug-in implementation leads to a rate-optimal choice of K. Chang et al. (2021) call this feature of the AR(1) plug-in implementation high-frequency compatible.

It should be noted that in the discrete-time setting it is typical to truncate the AR estimator at 0.97. See footnote 8 of Andrews (1991). Here, we should not follow this practice, as we rely on the convergence of $1-\hat{\rho}_j$ to zero at the rate of δ to achieve the high-frequency compatibility. Had we truncated the initial AR estimator at 0.97 or any fixed number less than 1, $\hat{\kappa}_D$ would be bounded away from zero with probability approaching one. As a result, $\hat{K}_D$ would be of order $n^{4/5}$ and we would lose the high-frequency compatibility. Computationally, without truncating the initial AR estimator, we may have $1-\hat{\rho}_j=0$ and encounter the "divided by zero" problem. To avoid this, we can truncate the AR estimator so that $1-\hat{\rho}_j$ is larger than the machine epsilon. In practice, $\{u_i\}_{i=1}^n$ is of course not observed, so $\hat{K}_D$ in (3.30) is computed utilizing $\{\hat{z}_i=x_i\hat{u}_i\}_{i=1}^n$ where $\hat{u}_i=y_i-x_i'\hat{\beta}_D$.

Note that the high-frequency compatible rate of K is of order $T^{4/5}$, which is smaller than $n^{4/5}$ by an order of magnitude. So, when T is small, K may be small too, and the fixed-K asymptotics may be more accurate.

The above MSE-optimal choice of K is obtained under the rate assumption that $K\to\infty$ but at a slower rate than T. The so-obtained choice rule in (3.28) satisfies the rate assumption. One may wonder whether we can obtain an MSE-optimal choice of K under the "fixed-K" assumption that K is held fixed. The answer is no. Under the fixed-K asymptotics, Theorem 3.B.1 shows that the variance of $\hat{\Omega}^*$ is proportional to $1/K$ and the squared-bias is proportional to $1/T^2$. To minimize the dominating terms in the MSE, we would make K as large as possible. Such an approach would then drive K to infinity and make it incompatible with the "fixed-K" assumption to begin with. As an example, consider the case when $d=1$ and the Fourier basis functions in (3.3) are used. By Theorem 3.B.1 (i) and (iii), the dominating terms in the MSE are

$$\frac{1}{T^2}B_1+\frac{2}{K}\Omega^2,$$

as $c_{\phi,1}(K)=\frac{1}{2}\frac{1}{K}\sum_{j=1}^K\left[\phi_j^2(1)+\phi_j^2(0)\right]=1$. It is now clear that there is no fixed-value of K that minimizes the above: any fixed value of K is dominated by a larger value.

The above analysis shows that only the large-K asymptotic framework is theoretically coherent with an asymptotically optimal choice of K. Such an optimal choice of K is seemingly incompatible with the distributional approximation obtained under the fixed-K asymptotic theory. This is not the case, and we provide a justification here. Let $\mathscr{C}_\alpha(p,K)$ be the $(1-\alpha)$-quantile of the fixed-K asymptotic distribution of F_T, that is

$$\Pr\left(\frac{K}{K-p+1}F_{p,K-p+1} > \mathscr{C}_\alpha(p,K)\right) = \alpha.$$

Note that $K/(K+p-1)F_{p,K-p+1} \Rightarrow \chi_p^2/p$ as $K \to \infty$. Letting $K \to \infty$ in the above equation and using the dominated convergence theorem, we obtain

$$\Pr\left(\chi_p^2/p > \lim_{K\to\infty}\mathscr{C}_\alpha(p,K)\right) = \alpha.$$

This shows that $\lim_{K\to\infty}\mathscr{C}_\alpha(p,K) = \chi_{p,\alpha}^2/p$, where $\chi_{p,\alpha}^2$ is the $(1-\alpha)$-quantile of the chi-squared distribution χ_p^2. Therefore, under the large K asymptotics, $\mathscr{C}_\alpha(p,K)$ is an asymptotically valid critical value, even though it is based on the fixed-K asymptotic distribution. In the literature on the fixed-smoothing asymptotics for discrete-time data with a fixed δ, it has been proved that for the location models and linear regression models, critical values based on the fixed-smoothing asymptotic distribution (i.e., K is fixed for series LRV estimation) are second-order correct under the increasing-smoothing asymptotics (i.e., $K \to \infty$). See, for example, Sun (2013) for the case with series LRV estimation and Sun (2014a) and Sun et al. (2008) for the case with kernel LRV estimation.

3.B.2 Proof of Theorem 3.B.1

Part (a): For notational simplicity, we assume that $\Omega^{1/2}$ is symmetric. Note that

$$\begin{aligned}
&\operatorname{var}\left[\operatorname{vec}\left(\Omega^{1/2}\frac{1}{K}\sum_{j=1}^{K}\left[\int_0^1 \phi_j(r)\,dW_d(r)\right]^{\otimes 2}\Omega^{1/2}\right)\right] \\
&=\frac{1}{K^2}\operatorname{var}\left[\left(\Omega^{1/2}\otimes\Omega^{1/2}\right)\operatorname{vec}\left(\sum_{j=1}^{K}\left[\int_0^1 \phi_j(r)\,dW_d(r)\right]^{\otimes 2}\right)\right] \\
&=\frac{1}{K^2}\left(\Omega^{1/2}\otimes\Omega^{1/2}\right)\operatorname{var}\left[\operatorname{vec}\left(\sum_{j=1}^{K}\left[\int_0^1 \phi_j(r)\,dW_d(r)\right]^{\otimes 2}\right)\right]\left(\Omega^{1/2}\otimes\Omega^{1/2}\right) \\
&=\frac{1}{K}\left(\Omega^{1/2}\otimes\Omega^{1/2}\right)\left(\mathbb{I}_{d^2}+\mathbb{K}_{dd}\right)\left(\Omega^{1/2}\otimes\Omega^{1/2}\right) \\
&=\frac{1}{K}\left(\Omega^{1/2}\otimes\Omega^{1/2}\right)\left(\Omega^{1/2}\otimes\Omega^{1/2}\right)\left(\mathbb{I}_{d^2}+\mathbb{K}_{dd}\right) \\
&=\frac{1}{K}\left(\Omega\otimes\Omega\right)\left(\mathbb{I}_{d^2}+\mathbb{K}_{dd}\right).
\end{aligned}$$

Hence, under Assumption 3.B.1(i), we have

$$\operatorname{var}\left[\operatorname{vec}(\hat{\Omega}^*)\right]=\frac{1}{K}\left(\Omega\otimes\Omega\right)\left(\mathbb{I}_{d^2}+\mathbb{K}_{dd}\right)\left(1+o(1)\right).$$

Part (b): Before computing the bias when $T\to\infty$ and $K\to\infty$, we first show that $\delta\sum_{k=-\infty}^{\infty}|k\delta|^m\left\|\Gamma_{XU}(k\delta)\right\|<\infty$ for $m=0,1,2$. Using Assumption 3.B.1(ii), we have

$$\begin{aligned}
\delta\sum_{|k|<n}|k\delta|^m\left\|\Gamma_{XU}(k\delta)\right\| &= \delta\sum_{|k\delta|\le C_2}|k\delta|^m\left\|\Gamma_{XU}(k\delta)\right\|+\delta\sum_{C_2<|k\delta|<n}|k\delta|^m\left\|\Gamma_{XU}(k\delta)\right\| \\
&\le \delta\sum_{|k\delta|\le C_2}C_2^m C_1+C_1\delta\sum_{C_2<|k\delta|<n}|k\delta|^m(k\delta)^{-(3+\iota)} \\
&= 2C_2^m C_1\delta\cdot\frac{C_2}{\delta}+C_1\delta^{m-2-\varepsilon}\sum_{C_2<|k\delta|<n}|k|^{-(3-m+\iota)} \\
&\le 2C_2^{m+1}C_1+C_1\delta^{m-2-\varepsilon}\cdot O\left(\frac{C_2}{\delta}\right)^{1-(3-m+\iota)} \\
&= 2C_2^{m+1}C_1+C_1\cdot O\left(C_2^{1-(3-m+\iota)}\right)=O(1).
\end{aligned}$$

So we have $\delta \sum_{k=-\infty}^{\infty} |k\delta|^m \|\Gamma_{XU}(k\delta)\| < \infty$. By the same argument, Assumption 3.B.1(ii) implies that $\int_{-\infty}^{\infty} |\tau|^m \|\Gamma_{XU}(\tau)\| < \infty$ for $m = 0, 1, 2$.

Next, we compute the bias of $\hat{\Omega}^*$ when $T \to \infty$ and $K \to \infty$. Denote $E\left[(x_i u_i)(x_\ell u_\ell)'\right] = \Gamma_{xu}(i-\ell)$. Note that

$$\begin{aligned}
&E(\hat{\Omega}^*) \\
&= \frac{1}{K}\sum_{j=1}^{K}\left[\frac{1}{\Lambda(n,\delta)^2}\sum_{i=1}^{n}\sum_{\ell=1}^{n}\phi_j\left(\frac{i}{n}\right)\phi_j\left(\frac{\ell}{n}\right)E\left(x_i u_i\right)\left(x_\ell u_\ell\right)'\right] \\
&= \frac{1}{K}\sum_{j=1}^{K}\frac{1}{\Lambda(n,\delta)^2}\sum_{i=1}^{n}\sum_{\ell=1}^{n}\phi_j\left(\frac{i}{n}\right)\phi_j\left(\frac{\ell}{n}\right)\Gamma_{xu}(i-\ell) \\
&= \frac{1}{K}\sum_{j=1}^{K}\frac{1}{\Lambda(n,\delta)^2}\sum_{i=1}^{n}\sum_{k=i-n}^{i-1}\phi_j\left(\frac{i}{n}\right)\phi_j\left(\frac{i-k}{n}\right)\Gamma_{xu}(k) \\
&= \frac{1}{K}\sum_{j=1}^{K}\frac{n}{\Lambda(n,\delta)^2}\sum_{k=-n+1}^{n-1}\left\{\frac{1}{n}\sum_{i=1}^{n}1\left\{\frac{1}{n}\le\frac{i-k}{n}\le 1\right\}\phi_j\left(\frac{i}{n}\right)\phi_j\left(\frac{i-k}{n}\right)\right\}\Gamma_{xu}(k) \\
&= \frac{1}{K}\sum_{j=1}^{K}\delta\sum_{k=-n+1}^{n-1}\omega_{j,n}\left(\frac{k}{n}\right)\Gamma_{xu}(k)
\end{aligned}$$

where

$$\omega_{j,n}\left(\frac{k}{n}\right) = \frac{1}{n}\sum_{i=1}^{n}1\left\{\frac{1}{n}\le\frac{i-k}{n}\le 1\right\}\phi_j\left(\frac{i}{n}\right)\phi_j\left(\frac{i-k}{n}\right).$$

The bias is then equal to

$$\begin{aligned}
\mathscr{B}_n &= E\hat{\Omega}^* - \Omega \\
&= \frac{1}{K}\sum_{j=1}^{K}\delta\sum_{k=-n+1}^{n-1}\left[\omega_{j,n}\left(\frac{k}{n}\right)-1\right]\Gamma_{xu}(k) + \delta\sum_{k=-n+1}^{n-1}\Gamma_{xu}(k) - \Omega \\
&:= \mathscr{B}_{1n} + \mathscr{B}_{2n}.
\end{aligned}$$

For $\mathscr{B}_{2n}$, we use Assumption 3.B.1(iii) with $m=0$ to obtain:

$$\begin{aligned}
\mathscr{B}_{2n} &= \delta \sum_{k=-n+1}^{n-1} \Gamma_{xu}(k) - \Omega = \delta \sum_{k=-n+1}^{n-1} \Gamma_{XU}(k\delta) - \Omega \\
&= \delta \sum_{k=-n+1}^{n-1} \Gamma_{XU}(k\delta) - \int_{-T}^{T} \Gamma_{XU}(\tau)\, d\tau + O\left(\frac{1}{T^2}\right) \\
&= O(\delta) + O\left(\frac{1}{T^2}\right),
\end{aligned}$$

where the $O\left(T^{-2}\right)$ term holds because under Assumption 3.B.1(ii),

$$\begin{aligned}
&\left\| \int_{-\infty}^{\infty} \Gamma_{XU}(\tau)\, d\tau - \int_{-T}^{T} \Gamma_{XU}(\tau)\, d\tau \right\| \\
&= \left\| \int_{-\infty}^{\infty} 1\{|\tau| \geq T\}\, \Gamma_{XU}(\tau)\, d\tau \right\| \leq \frac{1}{T^2} \int_{-\infty}^{\infty} \tau^2 1\{|\tau| \geq T\} \left\|\Gamma_{XU}(\tau)\right\| d\tau \\
&\leq \frac{1}{T^2} \int_{-\infty}^{\infty} \tau^2 \left\|\Gamma_{XU}(\tau)\right\| d\tau = O\left(\frac{1}{T^2}\right).
\end{aligned}$$

For $\mathscr{B}_{1n}$, we have, using $\sup_{r\in[0,1]} \left|\dot{\phi}_j(r)\right| \leq jC$ in Assumption 3.B.1(iv):

$$\begin{aligned}
\omega_{j,n}(\varsigma) &= \frac{1}{n} \sum_{i=1}^{n} 1\left\{\frac{1}{n} \leq \frac{i}{n} - \varsigma \leq 1\right\} \phi_j\left(\frac{i}{n}\right) \phi_j\left(\frac{i}{n} - \varsigma\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} 1\left\{\frac{1}{n} + \varsigma \leq \frac{i}{n} \leq 1 + \varsigma\right\} \phi_j\left(\frac{i}{n}\right) \phi_j\left(\frac{i}{n} - \varsigma\right) \\
&= \int_{\max(0,\varsigma)}^{\min(1+\varsigma,1)} \phi_j(r)\, \phi_j(r-\varsigma)\, dr + O\left(\frac{j}{n}\right) \\
&:= \omega_j(\varsigma) + O\left(\frac{j}{n}\right),
\end{aligned}$$

uniformly over $j = 1, 2, \ldots, K$ and $\varsigma \in [-1,1]$ where

$$\omega_j(\varsigma) = \int_{\max(0,\varsigma)}^{\min(1+\varsigma,1)} \phi_j(r)\, \phi_j(r-\varsigma)\, dr.$$

Note that $\omega_j(0) = 1$. Then we have, as $n \to \infty$,

$$\begin{aligned}
\mathscr{B}_{1n} &= \frac{1}{K}\sum_{j=1}^{K}\delta\sum_{k=-n+1}^{n-1}\left[\omega_{j,n}\left(\frac{k}{n}\right)-1\right]\Gamma_{xu}(k)\\
&= \frac{1}{K}\sum_{j=1}^{K}\delta\sum_{k=-n+1}^{n-1}\left[\omega_{j}\left(\frac{k}{n}\right)-1\right]\Gamma_{xu}(k)+\delta\sum_{k=-n+1}^{n-1}\left[\frac{1}{K}\sum_{j=1}^{K}O\left(\frac{j}{n}\right)\right]\Gamma_{xu}(k)\\
&= \frac{1}{K}\sum_{j=1}^{K}\delta\sum_{k=-n+1}^{n-1}\left[\omega_{j}\left(\frac{k}{n}\right)-1\right]\Gamma_{xu}(k)+O\left(\frac{K}{n}\right)\delta\sum_{k=-n+1}^{n-1}\|\Gamma_{XU}(k\delta)\|\\
&= \frac{1}{K}\sum_{j=1}^{K}\delta\sum_{k=-n+1}^{n-1}\left[\omega_{j}\left(\frac{k}{n}\right)-1\right]\Gamma_{xu}(k)+O\left(\frac{K}{n}\right)\\
&:= \tilde{\mathscr{B}}_{1n}+O\left(\frac{K}{n}\right),
\end{aligned}$$

where $\tilde{\mathscr{B}}_{1n} = \frac{1}{K}\sum_{j=1}^{K}\delta\sum_{k=-n+1}^{n-1}\left[\omega_j\left(\frac{k}{n}\right)-1\right]\Gamma_{xu}(k)$.

Now,

$$\begin{aligned}
\tilde{\mathscr{B}}_{1n} &= \frac{1}{K}\sum_{j=1}^{K}\delta\sum_{k=-n+1}^{n-1}\left[\omega_{j}\left(\frac{k}{n}\right)-1\right]\Gamma_{xu}(k)\\
&= \delta\sum_{n/\log n<|k|\le n-1}\left[\frac{1}{K}\sum_{j=1}^{K}\omega_j\left(\frac{k}{n}\right)-1\right]\Gamma_{xu}(k)+\delta\sum_{|k|\le n/\log n}\left[\frac{1}{K}\sum_{j=1}^{K}\omega_j\left(\frac{k}{n}\right)-1\right]\Gamma_{xu}(k)\\
&= \tilde{\mathscr{B}}_{11,n}+\tilde{\mathscr{B}}_{12,n}
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\mathscr{B}}_{11,n} &= \delta\sum_{n/\log n<|k|\le n-1}\left[\frac{1}{K}\sum_{j=1}^{K}\omega_j\left(\frac{k}{n}\right)-1\right]\Gamma_{xu}(k)\\
&\le \delta\sum_{n/\log n<|k|\le n-1}\left|\frac{1}{K}\sum_{j=1}^{K}\omega_j\left(\frac{k}{n}\right)-1\right|\left(\frac{k}{n/\log n}\right)^2\|\Gamma_{XU}(k\delta)\|\\
&= C\left(\frac{\log n}{n}\right)^2\frac{1}{\delta^2}\left[\delta\sum_{k=-\infty}^{\infty}(k\delta)^2\|\Gamma_{XU}(k\delta)\|\right] = O\left(\frac{(\log n)^2}{T^2}\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathscr{B}_{12,n} &= \frac{1}{K}\sum_{j=1}^{K}\delta\sum_{|k|\leq n/\log n}\left[\omega_j\left(\frac{k}{n}\right)-1\right]\Gamma_{xu}(k)\\
&= \frac{1}{K}\sum_{j=1}^{K}\delta\sum_{|k|\leq n/\log n}\left[\omega_j\left(\frac{k}{n}\right)-1\right]\Gamma_{XU}(k\delta)\\
&= \frac{1}{K}\sum_{j=1}^{K}\delta\sum_{|k|\leq n/\log n}\left[\dot{\omega}_j(0)\frac{k}{n}+\frac{1}{2}\ddot{\omega}_j\left(\frac{\tilde{k}}{n}\right)\left(\frac{k}{n}\right)^2\right]\Gamma_{XU}(k\delta)\\
&= \frac{1}{n\delta}\left[\frac{1}{K}\sum_{j=1}^{K}\dot{\omega}_j(0)\right]\delta\sum_{|k|\leq n/\log n}k\delta\Gamma_{XU}(k\delta)\\
&+\left(\frac{1}{n\delta}\right)^2\frac{1}{K}\sum_{j=1}^{K}\delta\sum_{|k|\leq n/\log n}\left[\frac{1}{2}\ddot{\omega}_j\left(\frac{\tilde{k}}{n}\right)\right](k\delta)^2\Gamma_{XU}(k\delta)\\
&= \frac{K^2}{T^2}\frac{1}{K^3}\sum_{j=1}^{K}\frac{1}{2}\ddot{\omega}_j(0)\,\delta\sum_{|k|\leq n/\log n}(k\delta)^2\Gamma_{XU}(k\delta)(1+o(1))+O\left(\frac{1}{n\delta}\frac{1}{K}\sum_{j=1}^{K}\dot{\omega}_j(0)\right)\\
&= \frac{K^2}{T^2}\left(\frac{1}{K^3}\sum_{j=1}^{K}\frac{1}{2}\ddot{\omega}_j(0)\right)\left(\int_{-\infty}^{\infty}\tau^2\Gamma_{XU}(\tau)\,d\tau\right)(1+o(1))+O\left(\frac{1}{n\delta}\frac{1}{K}\sum_{j=1}^{K}\dot{\omega}_j(0)\right).
\end{aligned}$$

Given that $\omega_j(\varsigma)=\int_{\varsigma}^{1}\phi_j(r)\,\phi_j(r-\varsigma)\,dr$, we have

$$\begin{aligned}
\dot{\omega}_j(\varsigma) &= -\phi_j(\varsigma)\,\phi_j(0)-\int_{\varsigma}^{1}\phi_j(r)\,\dot{\phi}_j(r-\varsigma)\,dr,\\
\ddot{\omega}_j(\varsigma) &= -\dot{\phi}_j(\varsigma)\,\phi_j(0)+\phi_j(\varsigma)\,\dot{\phi}_j(0)+\int_{\varsigma}^{1}\phi_j(r)\,\ddot{\phi}_j(r-\varsigma)\,dr=\int_{\varsigma}^{1}\phi_j(r)\,\ddot{\phi}_j(r-\varsigma)\,dr.
\end{aligned}$$

So

$$\begin{aligned}
\dot{\omega}_j(0) &= -\phi_j^2(0)-\frac{1}{2}\left[\phi_j^2(1)-\phi_j^2(0)\right]=-\frac{1}{2}\left[\phi_j^2(1)+\phi_j^2(0)\right],\\
\ddot{\omega}_j(0) &= \int_{0}^{1}\phi_j(r)\,\ddot{\phi}_j(r)\,dr.
\end{aligned}$$

Therefore, under Assumptions 3.B.1(iv) and (v), we have

$$\begin{aligned}
\tilde{\mathscr{B}}_{12,n} &= \frac{K^2}{T^2}\frac{1}{K^3}\sum_{j=1}^{K}\frac{1}{2}\ddot{\omega}_j(0)\,\delta\sum_{|k|\le n/\log n}(k\delta)^2\,\Gamma_{XU}(k\delta) \\
&= \frac{K^2}{T^2}\left(\frac{1}{K^3}\sum_{j=1}^{K}\frac{1}{2}\int_0^1\phi_j(r)\,\ddot{\phi}_j(r)\,dr\right)\int_{-\infty}^{\infty}\tau^2\Gamma_{XU}(\tau)\,d\tau\,(1+o(1))+O\left(\frac{1}{T}\right) \\
&= -\frac{K^2}{T^2}c_\phi\int_{-\infty}^{\infty}\tau^2\Gamma_{XU}(\tau)\,d\tau\,(1+o(1))+O\left(\frac{1}{T}\right)
\end{aligned}$$

as $K\to\infty$ and $T\to\infty$.

Combining the above results yields the asymptotic bias formula for the case where $K\to\infty$ and $T\to\infty$.

Part (c): As in the proof of Part (b), we have

$$\mathscr{B}_n = E(\hat{\Omega}^*)-\Omega := \tilde{B}_{12,n}+O\left(\frac{K}{n}+\delta+\frac{1}{T^2}+\frac{(\log n)^2}{T^2}\right),$$

where

$$\tilde{\mathscr{B}}_{12,n} = \frac{1}{K}\sum_{j=1}^{K}\delta\sum_{|k|\le n/\log n}\left[\omega_j\left(\frac{k}{n}\right)-1\right]\Gamma_{XU}(k\delta).$$

For the rest of the proof, we use arguments different from that for Part (b). Using $\omega_j(0)=1$ and $\dot{\omega}_j(0)=-\frac{1}{2}\left[\phi_j^2(1)+\phi_j^2(0)\right]$, we have

$$\begin{aligned}
\tilde{\mathscr{B}}_{12,n} &= \frac{1}{K}\sum_{j=1}^{K}\delta\sum_{|k|\le n/\log n}\left[\omega_j\left(\frac{k}{n}\right)-1\right]\Gamma_{XU}(k\delta) \\
&= \frac{1}{K}\sum_{j=1}^{K}\frac{1}{n}\delta\sum_{|k|\le n/\log n}\left[\dot{\omega}_j\left(\frac{\tilde{k}}{n}\right)k\right]\Gamma_{XU}(k\delta) \\
&= \frac{1}{K}\sum_{j=1}^{K}\frac{1}{n}\frac{1}{\delta}\dot{\omega}_j(0)\left[\delta\sum_{k=-\infty}^{\infty}[k\delta]\,\Gamma_{XU}(k\delta)+o(1)\right] \\
&= -\frac{1}{2}\frac{1}{T}\left(\frac{1}{K}\sum_{j=1}^{K}\left[\phi_j^2(1)+\phi_j^2(0)\right]\right)\int_{-\infty}^{\infty}\tau\Gamma_{XU}(\tau)\,d\tau\,(1+o(1)).
\end{aligned}$$

Therefore,

$$\mathscr{B}_n = -\frac{1}{2}\frac{1}{T}\left(\frac{1}{K}\sum_{j=1}^{K}\left[\phi_j^2(1)+\phi_j^2(0)\right]\right)\int_{-\infty}^{\infty}\tau\Gamma_{XU}(\tau)\,d\tau$$
$$+o\left(\frac{1}{T}\right)+O\left(\frac{1}{n}+\delta+\frac{(\log n)^2}{T^2}\right).$$

3.B.3 Sufficient conditions for Assumption 3.B.1(i)

We now provide sufficient conditions for Assumption 3.B.1(i). For notational simplicity, we consider the case that $v_i = x_i u_i$ is a scalar. The vector case requires only additional matrix algebra. The underlying continuous time process is $V_t = X_t U_t$. Let $v^* = (v_1^*,...,v_n^*)$ be a zero-mean Gaussian sequence with the same covariance as $v = (v_1,...,v_n)$. Then the fourth-order cumulant $\kappa_{v,4}(\ell_1,\ell_2,\ell_3,\ell_4)$ of $\{v_i\}_{i=1}^n$ is defined to be

$$\kappa_{v,4}(\ell_1,\ell_2,\ell_3,\ell_4) = E\left(v_{\ell_1}v_{\ell_1+\ell_2}v_{\ell_1+\ell_3}v_{\ell_1+\ell_4}\right) - E\left(v_{\ell_1}^*v_{\ell_1+\ell_2}^*v_{\ell_1+\ell_3}^*v_{\ell_1+\ell_4}^*\right).$$

We need the following assumption.

Assumption 3.B.2 (i) *v_i is fourth-order stationary with covariance $\Gamma_v(k) = E(v_i v_{i-k})$ and fourth-order cumulant $\kappa_{v,4}(\ell_1,\ell_2,\ell_3,\ell_4)$;* (ii) *there is a constant C that does not depend on δ or n such that*

$$\delta^3\sum_{\ell_1=-n+1}^{n-1}\sum_{\ell_2=-n+1}^{n-1}\sum_{\ell_3=-n+1}^{n-1}\left|\kappa_{v,4}(0,\ell_1,\ell_2,\ell_3)\right| < C.$$

Assumption 3.B.2(ii) is the discrete analogue of its continuous counterpart

$$\int_{-T}^{T}\int_{-T}^{T}\int_{-T}^{T}\kappa_{V,4}(0,r_1,r_2,r_3)\,dr_1dr_2dr_3 < \infty,$$

where $\kappa_{V,4}$ is the fourth order cumulant of $\{V_t\}$. The above condition is the same as Assumption 2.3(b) in Lu and Park (2019).

Proposition 3.B.1 *Let Assumptions 3.2.2, 3.B.1(ii)-(iv), and 3.B.2 hold. If $K^2 = o(n)$ and $K = o(T)$, then as $\delta \to 0$ and $T \to \infty$,*

$$var(\hat{\Omega}^*) = var\left\{ \frac{1}{K} \sum_{j=1}^{K} \left[\frac{1}{\Lambda(n,\delta)} \sum_{i=1}^{n} \phi_j \left(\frac{i}{n} \right) v_i \right]^{\otimes 2} \right\} = \frac{1}{K} 2\Omega^2 (1 + o(1))$$

for both a fixed K and a growing K $(i.e., K \to \infty)$, that is, Assumption 3.B.1(i) holds.

Proof of Proposition 3.B.1. In the following, we write $\sum_{j_1=1}^{K} \sum_{j_2=1}^{K}$ as $\sum_{j_1,j_2}$ when there is no possibility of confusion. All results in this proof hold for both a fixed K and large K unless stated otherwise. We have

$$\begin{aligned}
& var(\hat{\Omega}^*) = var\left\{ \frac{1}{K} \sum_{j=1}^{K} \left[\frac{1}{\Lambda(n,\delta)} \sum_{i=1}^{n} \phi_j \left(\frac{i}{n} \right) v_i \right]^2 \right\} \\
& = \frac{1}{K^2 \Lambda(n,\delta)^2} \sum_{j_1,j_2} \sum_{i_1,i_2,i_3,i_4}^{n} \phi_{j_1}(\frac{i_1}{n}) \phi_{j_1}(\frac{i_2}{n}) \phi_{j_2}(\frac{i_3}{n}) \phi_{j_2}(\frac{i_4}{n}) E\left[(v_{i_1} v_{i_2} - E v_{i_1} v_{i_2}) (v_{i_3} v_{i_4} - E v_{i_3} v_{i_4}) \right] \\
& = \frac{1}{K^2 \Lambda(n,\delta)^2} \sum_{j_1,j_2} E \sum_{i_1=1}^{n} \sum_{k_1=i_1-n}^{i_1-1} \sum_{i_2=1}^{n} \sum_{k_2=i_2-n}^{i_2-1} \phi_{j_1}(\frac{i_1}{n}) \phi_{j_1}(\frac{i_1-k_1}{n}) \phi_{j_2}(\frac{i_2}{n}) \phi_{j_2}(\frac{i_2-k_2}{n}) \\
& \times (v_{i_1} v_{i_1-k_1} - E v_{i_1} v_{i_1-k_1}) (v_{i_2} v_{i_2-k_2} - E v_{i_2} v_{i_2-k_2}) .
\end{aligned}$$

Let

$$\phi_{j_1,j_2,j_3,j_4}(i_1,i_2,k_1,k_2) = \phi_{j_1}(\frac{i_1}{n}) \phi_{j_2}(\frac{i_2}{n}) \phi_{j_3}(\frac{k_1}{n}) \phi_{j_4}(\frac{k_2}{n}),\ \phi_{j_1,j_2,}(i_1,i_2) = \phi_{j_1}(\frac{i_1}{n}) \phi_{j_2}(\frac{i_2}{n}),$$

$$\mu_4(i_1,i_2,k_1,k_2) = E(v_{i_1} v_{i_1-k_1} v_{i_2} v_{i_2-k_2}),\ \mu_4^*(i_1,i_2,k_1,k_2) = E\left(v_{i_1}^* v_{i_1-k_1}^* v_{i_2}^* v_{i_2-k_2}^* \right).$$

Recall that $v^* = (v_1^*, ..., v_n^*)$ is a zero-mean Gaussian sequence with the same covariance as

$v=(v_1,...,v_n)$. We have

$$
\begin{aligned}
&\mu_4^*(i_1,i_2,k_1,k_2)\\
&:=E\left(v_{i_1}^*v_{i_1-k_1}^*v_{i_2}^*v_{i_2-k_2}^*\right)\\
&=E\left(v_{i_1}^*v_{i_1-k_1}^*\right)E\left(v_{i_2}^*v_{i_2-k_2}^*\right)+E\left(v_{i_1}^*v_{i_2}^*\right)E\left(v_{i_1-k_1}^*v_{i_2-k_2}^*\right)+E\left(v_{i_1}^*v_{i_2-k_2}^*\right)E\left(v_{i_1-k_1}^*v_{i_2}^*\right)\\
&=E(v_{i_1}v_{i_1-k_1})E(v_{i_2}v_{i_2-k_2})+E(v_{i_1}v_{i_2})E(v_{i_1-k_1}v_{i_2-k_2})+E(v_{i_1}v_{i_2-k_2})E(v_{i_1-k_1}v_{i_2}).
\end{aligned}
$$

By definition, $\mu_4(i_1,i_2,k_1,k_2)-\mu_4^*(i_1,i_2,k_1,k_2)=\kappa_{v,4}(i_1,-k_1,i_2-i_1,i_2-k_2-i_1)$. So

$$
\begin{aligned}
&var(\hat{\Omega}^*)\\
&=\frac{1}{K^2\Lambda(n,\delta)^4}\sum_{j_1,j_2}\sum_{i_1=1}^{n}\sum_{k=i_1-n}^{i_1-1}\sum_{i_2=1}^{n}\sum_{k=i_2-n}^{i_2-1}\phi_{j_1,j_1,j_2,j_2}(i_1,i_1-k_1,i_2,i_2-k_2)\,\kappa_{v,4}(i_1,-k_1,i_2-i_1,i_2-k_2-i_1)\\
&+\frac{1}{K^2\Lambda(n,\delta)^4}\sum_{j_1,j_2}\sum_{i_1=1}^{n}\sum_{k=i_1-n}^{i_1-1}\sum_{i_2=1}^{n}\sum_{k=i_2-n}^{i_2-1}\phi_{j_1,j_1,j_2,j_2}(i_1,i_1-k_1,i_2,i_2-k_2)E(v_{i_1}v_{i_2})E(v_{i_1-k_1}v_{i_2-k_2})\\
&+\frac{1}{K^2\Lambda(n,\delta)^4}\sum_{j_1,j_2}\sum_{i_1=1}^{n}\sum_{k=i_1-n}^{i_1-1}\sum_{i_2=1}^{n}\sum_{k=i_2-n}^{i_2-1}\phi_{j_1,j_1,j_2,j_2}(i_1,i_1-k_1,i_2,i_2-k_2)E(v_{i_1}v_{i_2-k_2})E(v_{i_1-k_1}v_{i_2})\\
&:=I_1+I_2+I_3.
\end{aligned}
$$

Using $\left|\phi_{j_1,j_1,j_2,j_2}(i_1,i_2,k_1,k_2)\right|\le C$ for some constant C, which holds under Assumption

3.2.2, we obtain

$$
\begin{aligned}
&|I_1| \\
&\leq \frac{1}{K^2\Lambda(n,\delta)^4} \sum_{j_1,j_2} \sum_{i_1=1}^{n} \sum_{k_1=i_1-n}^{i_1-1} \sum_{i_2=1}^{n} \sum_{k=i_2-n}^{i_2-1} \left|\phi_{j_1,j_1,j_2,j_2}(i_1,i_2,k_1,k_2)\right| \left|\kappa_{v,4}(i_1,-k_1,i_2-i_1,i_2-k_2-i_1)\right| \\
&\leq \frac{C}{K^2\Lambda(n,\delta)^4} \sum_{j_1,j_2} \sum_{i_1=1}^{n} \sum_{k_1=i_1-n}^{i_1-1} \sum_{i_2=1}^{n} \sum_{k=i_2-n}^{i_2-1} \left|\kappa_{v,4}(i_1,-k_1,i_2-i_1,i_2-k_2-i_1)\right| \\
&= \frac{C}{K^2\Lambda(n,\delta)^4} \sum_{j_1,j_2} \sum_{i_1=1}^{n} \sum_{k_1=i_1-n}^{i_1-1} \sum_{\ell_1=i_1-n}^{i_1-1} \sum_{k_2=i_1+\ell_1-n}^{i_1+\ell_1-1} \left|\kappa_{v,4}(0,-k_1-i_1,k_1-2i_1,i_2-k_2-2i_1)\right| \\
&\leq \frac{n/\delta^3}{\Lambda(n,\delta)^4} \frac{C}{K^2} \sum_{j_1,j_2} \left(\delta^3 \sum_{\ell_1=-n+1}^{n-1} \sum_{\ell_2=-n+1}^{n-1} \sum_{\ell_3=-n+1}^{n-1} \left|\kappa_{v,4}(0,\ell_1,\ell_2,\ell_3)\right| \right) \\
&= O\left(\frac{1}{T}\right),
\end{aligned}
$$

where we have used Assumption 3.B.2.

It remains to consider I_2 and I_3. Using change of variables repeatedly, we have

$$
\begin{aligned}
I_2 &= \frac{1}{K^2\Lambda(n,\delta)^4} \sum_{j_1,j_2} \sum_{i_1=1}^{n} \sum_{k_1=i_1-n}^{i_1-1} \sum_{i_2=1}^{n} \sum_{k_2=i_2-n}^{i_2-1} \phi_{j_1,j_1,j_2,j_2}(i_1,i_1-k_1,i_2,i_2-k_2) \\
&\times \Gamma_v(i_2-i_1)\Gamma_v(i_2-i_1-(k_2-k_1)) \\
&= \frac{1}{K^2\Lambda(n,\delta)^4} \sum_{j_1,j_2} \sum_{i_1=1}^{n} \sum_{k_1=i_1-n}^{i_1-1} \sum_{i_2=1}^{n} \sum_{k_2=i_2-n}^{i_2-1} \phi_{j_1,j_2}(i_1,i_2)\phi_{j_1,j_2}(i_1-k_1,i_2-k_2) \\
&\times \Gamma_v(i_2-i_1)\Gamma_v(i_2-i_1-(k_2-k_1)) \\
&= \frac{1}{K^2\Lambda(n,\delta)^4} \sum_{j_1,j_2} \sum_{i_1=1}^{n} \sum_{i=i_1-1}^{i_1-n} \sum_{k_1=i_1-n}^{i_1-1} \sum_{k_2=i_1-i-n}^{i_1-i-1} \phi_{j_1,j_2}(i_1,i_1-i)\phi_{j_1,j_2}(i_1-k_1,i_1-i-k_2) \\
&\times \Gamma_v(-i)\Gamma_v(-i-(k_2-k_1)) \\
&= \frac{1}{K^2\Lambda(n,\delta)^4} \sum_{j_1,j_2} \left\{ \sum_{i_1=1}^{n} \sum_{i=i_1-n}^{i_1-1} \phi_{j_1,j_2}(i_1,i_1-i)\Gamma_v(i) \right\}^2 \\
&= \frac{\delta^2}{K^2} \sum_{j_1,j_2} \left\{ \sum_{i=-n+1}^{n-1} \left[\frac{1}{n} \sum_{i_1=1}^{n} 1\left\{ \frac{1}{n} \leq \frac{i_1-i}{n} \leq 1 \right\} \phi_{j_1}\left(\frac{i_1}{n}\right) \phi_{j_2}\left(\frac{i_1-i}{n}\right) \right] \Gamma_v(i) \right\}^2.
\end{aligned}
$$

For any $\varsigma \in [0,1]$, define

$$\omega_{j_1,j_2,n}(\varsigma) = \frac{1}{n}\sum_{i_1=1}^{n} 1\left\{\frac{1}{n} \leq \frac{i_1}{n} - \varsigma \leq 1\right\}\phi_{j_1}\left(\frac{i_1}{n}\right)\phi_{j_2}\left(\frac{i_1}{n} - \varsigma\right).$$

Then

$$I_2 = \frac{\delta^2}{K^2}\sum_{j_1,j_2}\left[\sum_{i=-n+1}^{n-1}\omega_{j_1,j_2,n}\left(\frac{i}{n}\right)\Gamma_v(i)\right]^2.$$

Under Assumptions 3.2.2 and 3.B.1(iv), we have

$$\begin{aligned}
\omega_{j_1,j_2,n}(\varsigma) &= \frac{1}{n}\sum_{i_1=1}^{n} 1\left\{\frac{1}{n} + \varsigma \leq \frac{i_1}{n} \leq 1 + \varsigma\right\}\phi_{j_1}\left(\frac{i_1}{n}\right)\phi_{j_2}\left(\frac{i_1}{n} - \varsigma\right) \\
&= \int_{\max(0,\varsigma)}^{\min(1+\varsigma,1)}\phi_{j_1}(r)\phi_{j_2}(r-\varsigma)\,dr + O\left(\frac{\max(j_1,j_2)}{n}\right) \\
&:= \omega_{j_1,j_2}(\varsigma) + O\left(\frac{\max(j_1,j_2)}{n}\right),
\end{aligned}$$

uniformly over $j_1, j_2 \in [K]$. That is, there exists a constant C not dependent on j_1, j_2, ς, or K such that $\left|\omega_{j_1,j_2,n}(\varsigma) - \omega_{j_1,j_2,n}(\varsigma)\right| \leq C(j_1+j_2)/n$. We can choose C large enough so that $\sup_{j_1,j_2,\varsigma}\left|\omega_{j_1,j_2}(\varsigma)\right| \leq C$. Hence, for

$$I_{21} = \frac{\delta^2}{K^2}\sum_{j_1,j_2}\left[\sum_{i=-n+1}^{n-1}\omega_{j_1,j_2}\left(\frac{i}{n}\right)\Gamma_v(i)\right]^2,$$

we have

$$\begin{aligned}
I_2 &= I_{21} + O\left[\frac{\delta^2}{K^2}\sum_{j_1,j_2}\left(\frac{\max(j_1,j_2)}{n}\right)\sum_{i=-n+1}^{n-1}|\Gamma_v(i)|\right] \\
&= I_{21} + O\left[\frac{\delta}{K^2}\sum_{j_1,j_2}\left(\frac{\max(j_1,j_2)}{n}\right)\delta\sum_{i=-n+1}^{n-1}|\Gamma_v(i)|\right] \\
&= I_{21} + O\left(\frac{\delta}{K^2}\frac{K^3}{n}\right) = I_{21} + O\left(\frac{1}{K}\frac{TK^2}{n^2}\right) \\
&= I_{21} + o\left(\frac{1}{K}\right),
\end{aligned}$$

as $TK^2/n^2 = o(1)$. In the above, we have used $\delta \sum_{i=-n+1}^{n-1} |\Gamma_v(i)| < \infty$, which holds under Assumption 3.B.1(ii).

Note that under Assumptions 3.2.2 and 3.B.1(iv), we have

$$\begin{aligned}
&\omega_{j_1,j_2}(0) = 1\{j_1 = j_2\}, \\
&\dot{\omega}_{j_1,j_2}(\varsigma) = -\phi_{j_1}(\varsigma)\phi_{j_2}(0) - \int_{\varsigma}^{1} \phi_{j_1}(r)\dot{\phi}_{j_2}(r-\varsigma)\,dr,
\end{aligned}$$

where $\omega_{j_1,j_2}(0) = 1\{j_1 = j_2\}$, which follows from the orthonormality of $\{\phi_j\}$. So,

$$\sup_{j_1,j_2,\varsigma} \left|\dot{\omega}_{j_1,j_2}(\varsigma)\right| < Cj_2$$

for some constant $C > 0$.

Using the above expressions of the derivatives and taking a Taylor expansion, we have, for $i^*(i) \in [0,i]$,

$$\begin{aligned}
I_{21} &= \frac{\delta^2}{K^2}\sum_{j_1,j_2}\left[\sum_{i=-n+1}^{n-1} \omega_{j_1,j_2}\left(\frac{i}{n}\right)\Gamma_v(i)\right]^2 \\
&= \frac{\delta^2}{K^2}\sum_{j_1,j_2}\left[\sum_{i=-n+1}^{n-1} \Gamma_v(i)\,1\{j_1 = j_2\} + \sum_{i=-n+1}^{n-1} \dot{\omega}_{j_1,j_2}\left(\frac{i^*(i)}{n}\right)\frac{i}{n}\Gamma_v(i)\right]^2 \\
&= \frac{\delta^2}{K^2}\sum_{j_1,j_2}\left(\sum_{i=-n+1}^{n-1} \Gamma_v(i)\,1\{j_1 = j_2\}\right)^2 \\
&+ \frac{2}{K^2}\sum_{j_1=j_2}\left(\sum_{i=-n+1}^{n-1} \Gamma_v(i)\right)\left(\delta\sum_{i=-n+1}^{n-1} \dot{\omega}_{j_1,j_2}\left(\frac{i^*(i)}{n}\right)\frac{i}{n}\delta\Gamma_v(i)\right) \\
&+ \frac{\delta^2}{K^2}\sum_{j_1,j_2}\left[\sum_{i=-n+1}^{n-1} \dot{\omega}_{j_1,j_2}\left(\frac{i^*(i)}{n}\right)\frac{i}{n}\Gamma_v(i)\right]^2 \\
&= \frac{1}{K^2}\sum_{j=1}^{K}\left(\delta\sum_{i=-n+1}^{n-1} \Gamma_v(i)\right)^2 + O\left(\frac{1}{K^2}\frac{K^2}{n}\right) + O\left(\frac{1}{K}\frac{K^2}{T^2}\right) \\
&= \left(\delta\sum_{i=-n+1}^{n-1} \Gamma_v(i)\right)^2\frac{1}{K} + o\left(\frac{1}{K}\right),
\end{aligned}$$

where we have used $\delta \sum_{i=-n+1}^{n-1} |i| \, \delta \left| \Gamma_v(i) \right| < \infty$, which holds under Assumption 3.B.1(ii).

Now under Assumption 3.B.1(iii) with $m = 0$ and Assumption 3.B.1(ii), we have

$$\delta \sum_{i=-n+1}^{n-1} \Gamma_v(i) = \delta \sum_{i=-n+1}^{n-1} \Gamma_V(i\delta) \to \Omega.$$

Therefore, we have proved that $I_2 = \Omega/K\left(1 + o(1)\right)$. Similar arguments can be invoked to show that $I_3 = \Omega/K\left(1 + o(1)\right)$. Details are omitted here. Combining the results for I_1, I_2, and I_3 yields the desired result: $var(\hat{\Omega}^*) = 2\Omega^2/K\left(1 + o(1)\right)$. ■

Bibliography

Alquier, P., Ridgway, J., and Chopin, N. (2016). On the properties of variational approximations of gibbs posteriors. *The Journal of Machine Learning Research*, 17(1):8374–8414.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59:817–858.

Anthony, M. and Bartlett, P. L. (2009). *Neural network learning: Theoretical foundations*. cambridge university press.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Athey, S. and Wager, S. (2021). Policy learning with observational data. *Econometrica*, 89(1):133–161.

Babii, A., Chen, X., Ghysels, E., and Kumar, R. (2020). Binary choice with asymmetric loss in a data-rich environment: Theory and an application to racial justice.

Bartlett, P. L., Boucheron, S., and Lugosi, G. (2002). Model selection and error estimation. *Machine Learning*, 48(1-3):85–113.

Bégin, L., Germain, P., Laviolette, F., and Roy, J.-F. (2016). Pac-bayesian bounds based on the rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444.

Beygelzimer, A. and Langford, J. (2009). The offset tree for learning with partial labels. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 129–138.

Bhattacharya, D. and Dupas, P. (2012). Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167(1):168–196.

Bhattacharya, R. N. (1982). On the functional central limit theorem and the law of the iterated logarithm for markov processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 60(2):185–201.

Bilodeau, M. and Brenner, D. (2010). *Theory of Multivariate Statistics*. Springer-Verlag New York.

Bloom, H. S., Orr, L. L., Bell, S. H., Cave, G., Doolittle, F., Lin, W., and Bos, J. M. (1997). The benefits and costs of jtpa title ii-a programs: Key findings from the national job training partnership act study. *Journal of human resources*, pages 549–576.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

Catoni, O. (2004). *Statistical learning theory and stochastic optimization: Ecole d'Eté de Probabilités de Saint-Flour, XXXI-2001*, volume 1851. Springer Science & Business Media.

Catoni, O. (2007). Pac-bayesian supervised classification: The thermodynamics of statistical learning. institute of mathematical statistics lecture notes—monograph series 56. *IMS, Beachwood, OH. MR2483528*.

Chang, Y., Lu, Y., and Park, J. Y. (2021). Understanding regressions with observations collected at high frequency over long span. Working paper, Department of Economics, Indiana University.

Comte, F. (1999). Discrete and continuous time cointegration. *Journal of Econometrics*, 88(2):207–226.

Cover, T. M. and Thomas, J. A. (2006). Elements of information theory second edition solutions to problems. *Internet Access*, pages 19–20.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, pages 146–158.

Dehejia, R. H. (2005). Program evaluation as a decision problem. *Journal of Econometrics*, 125(1-2):141–173.

Del Moral, P., Doucet, A., and Jasra, A. (2006). Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436.

Donsker, M. D. and Varadhan, S. R. S. (1975). Asymptotic evaluation of certain markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28(1):1–47.

Elliott, G. and Lieli, R. P. (2013). Predicting binary outcomes. *Journal of Econometrics*, 174(1):15–26.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., and Group, O. H. S. (2012). The oregon health insurance experiment: Evidence from the first year. *The Quarterly Journal of Economics*, 127(3):1057–1106.

Fischer, M. and Nappo, G. (2009). On the moments of the modulus of continuity of itô processes. *Stochastic Analysis and Applications*, 28(1):103–122.

Freund, Y., Mansour, Y., and Schapire, R. E. (2004). Generalization bounds for averaged classifiers. *The Annals of Statistics*, 32(4):1698 – 1722.

Germain, P., Lacasse, A., Laviolette, F., March, M., and Roy, J.-F. (2015). Risk bounds for the majority vote: From a pac-bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16(26):787–860.

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. (2009). Pac-bayesian learning of linear classifiers. *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 353–360.

Granger, C. W. and Machina, M. J. (2006). Forecasting and decision theory. volume 1 of *Handbook of Economic Forecasting*, pages 81–98. Elsevier.

Guedj, B. (2013). *Aggregation of estimators and classifiers: theory and methods*. PhD thesis, Université Pierre et Marie Curie-Paris VI.

Guedj, B. and Robbiano, S. (2018). Pac-bayesian high dimensional bipartite ranking. *Journal of Statistical Planning and Inference*, 196:70–86.

Haussler, D. (1992). Decision theoretic generalizations of the pac model for neural net and other learning applications. *Information and computation*, 100(1):78–150.

Hendren, N. and Sprung-Keyser, B. (2020). A unified welfare analysis of government policies. *The Quarterly Journal of Economics*, 135(3):1209–1318.

Hirano, K. and Porter, J. R. (2009). Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701.

Horowitz, J. L. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica*, 60(3):505–531.

Huang, X. and Xu, J. (2020). Estimating individualized treatment rules with risk constraint. *Biometrics*, 76(4):1310–1318.

Hwang, J. and Sun, Y. (2018). Simple, robust, and accurate F and t tests in cointegrated systems. *Econometric Theory*, 34(5):949—-984.

Jasra, A., Stephens, D. A., and Holmes, C. C. (2007). On population-based simulation for static inference. *Statistics and Computing*, 17(3):263–279.

Jiang, W. and Tanner, M. A. (2008). Gibbs posterior for variable selection in high-dimensional classification and data mining. *The Annals of Statistics*, 36(5):2207 – 2231.

Jin, S., Phillips, P. C. B., and Sun, Y. (2006). A new approach to robust inference in cointegration. *Economics Letters*, 91(2):300 – 306.

Kaplan, D. M. and Sun, Y. (2017). Smoothed estimating equations for instrumental variables quantile regression. *Econometric Theory*, 33(1):105–157.

Kiefer, N. M. and Vogelsang, T. J. (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, 21(6):1130—1164.

Kim, J. and Park, J. Y. (2017). Asymptotics for recurrent diffusions with application to high frequency regression. *Journal of Econometrics*, 196(1):37–54.

Kitagawa, T., Lopez, H., and Rowley, J. (2023). Stochastic treatment choice with empirical welfare updating.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616.

Koltchinskii, V. (2001). Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914.

Kuecken, M., Thuilliez, J., and Valfort, M.-A. (2014). Does malaria control impact education? a study of the global fund in africa.

Lacasse, A., Laviolette, F., Marchand, M., Germain, P., and Usunier, N. (2006). PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. *Advances in Neural Information Processing Systems*, pages 769–776.

Langford, J. and Shawe-Taylor, J. (2003). Pac-bayes & margins. In *Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference*, volume 15, pages 439–446. MIT Press.

Lazarus, E., Lewis, D. J., Stock, J. H., and Watson, M. W. (2018). HAR inference: Recommendations for practice. *Journal of Business & Economic Statistics*, 36(4):541–559.

Lengeler, C. (1998). Insecticide treated bednets and curtains for malaria control. *Cochrane database of systematic reviews*, (2).

Lever, G., Laviolette, F., and Shawe-Taylor, J. (2010). Distribution-dependent pac-bayes priors. In *International Conference on Algorithmic Learning Theory*, pages 119–133. Springer.

London, B. and Sandler, T. (2019). Bayesian counterfactual risk minimization. In *International Conference on Machine Learning*, pages 4125–4133. PMLR.

Lu, Y. and Park, J. Y. (2019). Estimation of longrun variance of continuous time stochastic process using discrete sample. *Journal of Econometrics*, 210(2):236–267.

Manski, C. F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, 3(3):205–228.

Manski, C. F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics*, 27(3):313–333.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4):1221–1246.

Massart, P. (2007). *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer.

Massart, P. and Picard, J. (2007). *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*. Lecture Notes in Mathematics. Springer Berlin Heidelberg.

Maurer, A. (2004). A note on the pac bayesian theorem. *arXiv preprint cs/0411099*.

Mbakop, E. and Tabord-Meehan, M. (2021). Model selection for treatment choice: Penalized welfare maximization. *Econometrica*, 89(2):825–848.

McAllester, D. (2003a). Simplified pac-bayesian margin bounds. In *Learning theory and Kernel machines*, pages 203–215. Springer.

McAllester, D. A. (1999a). Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170.

McAllester, D. A. (1999b). Some pac-bayesian theorems. *Machine Learning*, 37(3):355–363.

McAllester, D. A. (2003b). Pac-bayesian stochastic model selection. *Machine Learning*, 51(1):5–21.

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in Combinatorics*, 141(1):148–188.

Newey, W. K. and West, K. D. (1994). Automatic lag selection in covariance matrix estimation. *The Review of Economic Studies*, 61(4):631–653.

Neyshabur, B., Bhojanapalli, S., Mcallester, D., and Srebro, N. (2017). Exploring generalization in deep learning. *Advances in Neural Information Processing Systems*, 30:5947–5956.

Orr, L. L., Lin, W., Cave, G., and Doolittle, F. (1994). *The national JTPA study: impacts, benefits, and costs of Title II-A*.

Pellatt, D. F. and Sun, Y. (2022). Asymptotic F test in regressions with observations collected at high frequency over long span. Working paper, Department of Economics, UC San Diego.

Phillips, P. C. B. (2005). HAC estimation by automated regression. *Econometric Theory*, 21(1):116–142.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ridgway, J., Alquier, P., Chopin, N., and Liang, F. (2014). Pac-bayesian auc classification and scoring. *arXiv preprint arXiv:1410.1771*.

Rozanov, Y. A. (1960). A central limit theorem for additive random functions. *Theory of Probability & Its Applications*, 5(2):221–223.

Seeger, M. (2002). Pac-bayesian generalisation error bounds for gaussian process classification. *Journal of Machine Learning Research*, 3(Oct):233–269.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shawe-Taylor, J. and Williamson, R. C. (1997). A pac analysis of a bayesian estimator. In *Proceedings of the tenth annual conference on Computational learning theory*, pages 2–9.

Stock, J. H. and Watson, M. W. (2019). *Introduction to Econometrics (Fourth Edition)*. Pearson.

Su, J.-H. (2020). Model selection in utility-maximizing binary prediction. *Journal of Econometrics*.

Sun, H., Du, S., and Wager, S. (2021). Treatment allocation under uncertain costs. *arXiv preprint arXiv:2103.11066*.

Sun, L. (2021). Empirical welfare maximization with constraints. *arXiv preprint arXiv:2103.15298*.

Sun, Y. (2004). A convergent t-statistic in spurious regressions. *Econometric Theory*, 20(5):943–962.

Sun, Y. (2011). Robust trend inference with series variance estimator and testing-optimal smoothing parameter. *Journal of Econometrics*, 164(2):345–366.

Sun, Y. (2013). A heteroskedasticity and autocorrelation robust F test using an orthonormal series variance estimator. *The Econometrics Journal*, 16(1):1–26.

Sun, Y. (2014a). Fixed-smoothing asymptotics in a two-step generalized method of moments framework. *Econometrica*, 82(6):2327–2370.

Sun, Y. (2014b). Let's fix it: Fixed-b asymptotics versus small-b asymptotics in heteroskedasticity and autocorrelation robust inference. *Journal of Econometrics*, 178:659–677.

Sun, Y. (2022). Some extensions of asymptotic F and t theory in nonstationary regressions. Technical report.

Sun, Y., Phillips, P. C. B., and Jin, S. (2008). Optimal bandwidth selection in heteroskedasticity–autocorrelation robust testing. *Econometrica*, 76(1):175–194.

Teklehaimanot, A., McCord, G. C., and Sachs, J. D. (2007). Scaling up malaria control in africa: An economic and epidemiological assessment. *The American Journal of Tropical Medicine and Hygiene*, 77(6 Suppl):138 – 144.

Thiemann, N., Igel, C., Wintenberger, O., and Seldin, Y. (2017). A strongly quasiconvex pac-bayesian bound. In Hanneke, S. and Reyzin, L., editors, *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, volume 76 of *Proceedings of Machine Learning Research*, pages 466–492. PMLR.

Tibshirani, J., Athey, S., Friedberg, R., Hadad, V., Hirshberg, D., Miner, L., Sverdrup, E., and Wager, S. (2022). *grf: Generalized Random Forests*. R package version 2.2.1.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142.

van der Vaart, A. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer Series in Statistics. Springer.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Vapnik, V. (1992). Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Viviano, D. (2019). Policy targeting under network interference. *arXiv preprint arXiv:1906.10258*.

Vogelsang, T. J. and Wagner, M. (2014). Integrated modified OLS estimation and fixed-b inference for cointegrating regressions. *Journal of Econometrics*, 178(2):741–760.

Wang, Y., Fu, H., and Zeng, D. (2018). Learning optimal personalized treatment rules in consideration of benefit and risk: with an application to treating type 2 diabetes patients with insulin therapies. *Journal of the American Statistical Association*, 113(521):1–13.

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118.