IDENTIFICATION AND INFERENCE FOR ALGORITHMIC FRONTIERS WITH SELECTIVE LABELS

YIQI LIU
Department of Economics, Cornell University

FRANCESCA MOLINARI
Department of Economics, Cornell University

AMILCAR VELEZ
Department of Economics, Cornell University

This paper provides identification results to characterize a fairness-accuracy (FA) frontier, and statistical inference tools to test hypotheses and build a confidence set for the FA-frontier, when outcomes are observed only for selected individuals. When the selection process is unrestricted but loss is measured in specific ways, we provide a characterization of the sharp identification region of the FA-frontier. Under an assumption of unconfoundedness conditional on observables (and unrestricted loss functions), we obtain point identification and propose a debiased machine learning estimator, derive its asymptotic distribution, and show how this can be used to carry out inference for the FA-frontier. In work in progress, we extend the partial identification results to a broader class of loss functions.

KEYWORDS: Algorithmic fairness, selective labels, statistical inference, support function.

1. INTRODUCTION

Over the last decade, algorithms have come to play an increasingly prevalent role in everyday life, often by supporting high-stake decisions through predictions of quantities

Yiqi Liu: yl3467@cornell.edu
Francesca Molinari: fm72@cornell.edu
Amilcar Velez: amilcare@cornell.edu
Preliminary version. This draft: March 26, 2026.

such as re-offense risk, repayment likelihood, and health care needs. These predictions enter institutional decision processes that determine who is detained or released, who is granted credit, and which patients receive outreach and care-management interventions. The increasing reliance on algorithms has prompted the emergence of a set of regulatory and oversight processes, known as *AI governance*, that rule the design, deployment, documentation, and monitoring of AI systems (including banning protected attributes as inputs to these systems) in order to manage risks to individuals and to ensure accountability. A distinctive feature of emerging regulatory regimes is their emphasis on evidence: entities deploying high-risk systems are increasingly expected to maintain traceability through record-keeping and logging, and to document and monitor accuracy and other performance metrics over the system's lifecycle (see, e.g., the EU AI Act, European Parliament and Council, 2024, Articles 11, 12 and 15). These requirements build on proposed documentation frameworks for datasets, models, and algorithmic impact assessments (Gebru et al., 2021, Mitchell et al., 2019, Raji et al., 2020).

Fairness and accuracy evaluations play a central role in governance, as they determine both the validity of model-supported decisions and whether those decisions impose disproportionate harms on legally protected groups.[1] At the same time, governance questions are rarely about a single fixed model. In public oversight and non-discrimination doctrine, the relevant question is often whether a challenged procedure can be replaced by an alternative that achieves comparable accuracy with less adverse impact, shifting attention from a single algorithm to a potentially large class of admissible decision rules, e.g., those obtained under constraints on inputs, model classes, or the use of sensitive attributes. Regulators and oversight bodies are called to adjudicate (i) whether a deployed rule is Pareto efficient, (ii) whether design restrictions enable less discriminatory alternatives (LDAs) with com-

[1] A growing literature documents that algorithms may exhibit systematic differences in performance across legally protected groups, both in predictive accuracy and in the decisions they induce (see, e.g., Angwin et al., 2016, Arnold et al., 2021, Obermeyer et al., 2019, Cowgill and Tucker, 2020, Berk et al., 2021). At the same time, improving parity in expected losses often requires sacrificing accuracy for at least one group, and conversely increasing accuracy for one group may worsen disparities.

parable accuracy, and (iii) whether such LDAs exist within an admissible class of decision rules.[2] We view the theoretical framework put forward by Liang, Lu, Mu, and Okumura (2025, LLMO henceforth) as well-suited to formalize these questions, becuase it characterizes a fairness-accuracy (FA) frontier under broad preferences over group-specific expected losses as well as the implications of banning attributes as inputs to the algorithms.

This paper extends the research agenda set out in Liu and Molinari (2026) to provide a toolkit for non-parametric estimation of and statistical inference on LLMO's FA-frontier and to assess the existence of less discriminatory alternatives (LDAs), when there are two groups of individuals and two possible decisions. The first extension we put forward is to allow for accuracy and fairness of an algorithm to be assessed through two different loss functions, thereby aligning our treatment with the literature in computer science (Corbett-Davies et al., 2017, Menon and Williamson, 2018, Kim et al., 2020, Gillis et al., 2024). We show that, as laid out in Liu and Molinari (2026)'s groundwork, the support function of a convex feasible set of expected accuracy losses for the two groups and the difference in expected fairness losses between the groups plays a crucial role in characterizing the FA-frontier and carrying out inference.

Our second extension recognizes that in many high-stakes settings, outcomes are observed only for selected individuals. For instance, pretrial misconduct (or lack thereof) is observed only if bail is granted; repayment behavior is observed only when credit is extended; and certain health outcomes are observed only for patients who receive treatment or remain under follow-up within a system. Hence, instead of observing the outcome vector of interest Y^*, one observes ZY^*, where $Z \in \{0, 1\}$ indicates observability. This *selective labels* problem creates challenges for governance compliance, similar to the ones long recognized in the vast econometrics literature on selectively observed data (e.g., Heckman, 1979, Manski, 1989). Accuracy metrics computed using only observed outcomes may mis-

[2]For example, US employment discrimination law prohibits selection procedures that cause disparate impact unless the employer demonstrates they are "job related for the position in question and consistent with business necessity." Even then the practice is unlawful if an alternative practice would serve the employer's legitimate interests with less discriminatory impact (U.S. Congress, 1991).

represent true performance, potentially overstating accuracy where it matters most (e.g., if those denied bail would not have re-offended). Hence, assessments of whether a system achieves "appropriate levels of accuracy" (EU AI Act, European Parliament and Council, 2024, Article 15) or whether LDAs exist (U.S. Congress, 1991) may be incorrect and required documentation may systematically exclude the most policy-relevant counterfactuals.

This paper studies the FA-properties of a family of decision rules $\mathcal{A}(\mathcal{X})$ (measurable functions mapping covariates $X \in \mathcal{X}$ to $[0,1]$) when label observability is selected. In the absence of restrictions on the selection process, we characterize the sharp identification region for the FA-frontier in the spirit of Manski (1989) (see, e.g., Molinari, 2020, Canay et al., 2026, for reviews of the partial identification literature), when classification loss is used to measure accuracy and statistical parity is used to measure fairness (as in, e.g., Menon and Williamson, 2018).

We then impose a missing-at-random (MAR) assumption, under which, after conditioning on X and a protected-group indicator G, observability does not convey additional information about Y^*. In this environment, and for any loss function, we provide a debiased machine learning (DML) estimator for the FA-frontier based on inverse probability weighting. Our DML approach is particularly appropriate for governance contexts, where high-dimensional administrative data enable flexible non-parametric modeling of selection, and formal inference with valid confidence statements remains necessary for legally defensible determinations. While ideas transfer from the program evaluation and high-dimensional inference literature (e.g., Chernozhukov et al., 2024, and references therein), we extend the methods to carry out inference on a new and set-valued estimand, the FA-frontier. We derive the asymptotic distribution of our estimator building on results in Liu and Molinari (2026), Semenova (2023), Chernozhukov et al. (2018), Beresteanu and Molinari (2008), and extend the testing methods in Liu and Molinari (2026), which build on Fang and Santos (2019), to account for inverse-probability weighting correction in the DML estimator. Our inference procedures yield confidence sets for the FA frontier and tests for efficiency and LDA existence. These directly support governance determinations: for instance, an en-

tity can demonstrate compliance with LDA requirements by showing the algorithm they use yields expected losses in the confidence set for the frontier.

Accompanying software implementing both the MAR-based and partial identification approaches is under development, along with empirical applications of our methods.

Related Literature. A large literature in computer science, statistics, and economics studies algorithmic fairness; see Chouldechova and Roth (2018), Barocas et al. (2023), and Corbett-Davies et al. (2024) for overviews. Much of this work treats fairness as a constraint or regularizer in an objective that prioritizes predictive performance (e.g., Dwork et al., 2012, Berk et al., 2017), or studies impossibility results and trade-offs among fairness criteria. A related empirical and methodological literature computes fairness-accuracy Pareto frontiers or ranges of disparities for specific model classes, typically as an optimization or auditing exercise rather than as a problem of econometric inference (e.g., Wei and Niethammer, 2022, Little et al., 2022, Coston et al., 2021). In contrast, the FA frontier of LLMO, for which we provide inference methods, is defined as the set of FA-undominated group loss pairs over a broad class of algorithms using a fixed information set, and it is designed to support policy questions about efficiency trade-offs and design restrictions.

Still within the algorithmic fairness literature, our work is related to studies concerning selective labels (e.g., Lakkaraju et al., 2017, Kleinberg et al., 2017, Rambachan et al., 2025, Khan et al., 2025). In particular, Rambachan et al. (2025) develop a general framework for robust evaluation and design of predictive algorithms under selective labels and unobserved confounding, delivering bounds for performance measures under alternative restrictions on the selection process. We view their contribution as addressing a distinct question from ours: how to evaluate predictive performance in the presence of counterfactual outcomes. Our focus is instead on identification and inference for the fairness-accuracy trade-offs of group losses across all feasible algorithms using a given information set, and on hypothesis tests and confidence statements about frontier properties such as efficiency and LDA existence that directly map to governance questions and design restrictions. For instance, our results can be used to characterize how excluding sensitive attributes or restricting the

model class affects the feasible set of group-loss pairs, or test whether a deployed algorithm is Pareto-dominated by alternatives within an admissible class.

Outline. Section 2 lays out notation and summarizes the derivation of the FA-frontier in Liu and Molinari (2026) and LLMO. Section 3 formalizes the selective labels problem, establishes point identification of the FA-frontier under a MAR assumption, puts forward our DML estimator, and derives its asymptotic distribution. Section 4 derives the asymptotic distribution under selectively observed labels of a statistic designed by Liu and Molinari (2026) to test for existence of an LDA. Section 5 concludes. Our main proofs are in Appendix A; Appendix B reports auxiliary results.

2. SETUP

Let a population of individuals be described by an outcome vector $Y^* \in \mathcal{Y} \subset \mathbb{R}^{d_Y}$, a binary group identity $G \in \{r, b\}$ (red or blue), and a vector of covariates $X \in \mathcal{X} \subset \mathbb{R}^{d_X}$. For example, one component of Y^* may denote an individual's number of active chronic illnesses in the subsequent year and the other component the cost of care; G may denote the individual's race, and X may include age, gender, biomarkers, comorbidity, and medication variables. As in Liu and Molinari (2026), we leave the relation between G and X unspecified, but require G not to be a deterministic function of X. Each individual receives a binary decision $D \in \{0, 1\}$, e.g., whether they are automatically enrolled in a high-risk care management program; or granted bail; or granted or denied a loan.[3] An algorithm $a : \mathcal{X} \mapsto [0, 1]$ assigns a probability distribution to D conditional on X; e.g., the algorithm assigns each patient a health risk score in $[0, 1]$; or a recidivism probability; or a repayment probability. For simplicity, we take $a(X)$ to be the only input to the decision, and hence $D \sim a(X)$. We let $\mathcal{A}(\mathcal{X})$ denote the set of all (measurable) algorithms that map from the input space $\mathcal{X}$ to a probability distribution over D. Extending the framework in Liu and Molinari (2026), we let $\ell^A : \{0, 1\} \times \mathcal{Y} \mapsto \mathbb{R}$ be a loss function used to measures the *accuracy* of an algorithm for an individual with outcome $y \in \mathcal{Y}$, and $\ell^F : \{0, 1\} \times \mathcal{Y} \mapsto \mathbb{R}$ a loss function used to measures the algorithm's *fairness*. For example, one could use classifica-

[3] While we assume G and D to be binary, the results can be extend to multiple groups and multiple decisions.

tion error to measure accuracy ($\ell^A(d,y) = \mathbb{1}(d \neq y_1)$), which in our health care example returns the value 1 if the algorithm mistakenly enrolls a healthy person in the high-risk care program or fails to enroll someone who is very sick. And one could use statistical parity to measure fairness ($\ell^F(d,y) = 1$ if $d = 1$ and $\ell^F(d,y) = 0$ otherwise), hence judging an algorithm more fair if the proportion of either group receiving the two decisions is closer.

This paper is concerned with the case where instead of observing Y^*, we observe $Y = ZY^*$, where $Z \in \{0,1\}$ is an indicator variable that records when the outcome of interest Y^* is observed in the testing data. In other words, we allow for selectively observed labels: in the case of health applications, this may be due to selection into treatment; for bail decision, it may be due to observing whether a defendant returned for their court appearance without committing a crime only for people let out on bail; for mortgage decisions, it may be due to observing repayment behavior only for people who are issued a mortgage.

Before analyzing this problem, we provide a tractable characterization, under the assumption that $(Y^*, X, G, Z) \sim \mathbb{P}^*$ is observed, of the FA-frontier for the case that the loss function used to measure accuracy differs from that used to measure fairness. In doing so, we extend the analysis in Appendix O.1 of LLMO. Armed with our novel characterization, we study what can be learned about the FA-frontier and how can valid statistical inference be carried out, when in fact only the selectively observed labels Y are available.

2.1. *The Fairness-Accuracy Frontier*

We denote $\mu_g \equiv \mathbb{P}^*(G = g)$ the population proportion of group $g \in \{r, b\}$ and, for $d \in \{0,1\}$, we denote the (unobserved) *ideal labels* associated with loss function $\iota \in \{A, F\}$ as

$$L_d^{g,\iota} \equiv \frac{\ell^\iota(d,Y^*)\mathbb{1}\{G=g\}}{\mu_g}. \tag{1}$$

We denote the conditional expectation of $L_d^{g,\iota}$ taken with respect to $\mathbb{P}^*(Y^*, G|X)$ by

$$\theta_d^{g,\iota}(X) \equiv \mathbb{E}^*[L_d^{g,\iota}|X] = \frac{\mathbb{E}^*[\ell^\iota(d,Y^*)\mathbb{1}\{G=g\}|X]}{\mu_g}.$$

DEFINITION 1—Ideal labels and conditional expectations: For $d \in \{0,1\}$, define

$$\mathbf{L}_d \equiv \begin{bmatrix} L_d^{r,A} \\ L_d^{b,A} \\ L_d^{r,F} - L_d^{b,F} \end{bmatrix} \in \mathbb{R}^3, \tag{2}$$

and let $\boldsymbol{\theta}_d(X) \equiv \mathbb{E}^*[\mathbf{L}_d|X] \in \mathbb{R}^3$ and $\Delta\boldsymbol{\theta}(X) \equiv \boldsymbol{\theta}_1(X) - \boldsymbol{\theta}_0(X) \in \mathbb{R}^3$.

To make sure that the ideal labels are well-defined, we assume:

ASSUMPTION 1—Moment restrictions for both losses: *There exist constants* $0 < c_1 < 1$ *and* $0 < c_2 < \infty$ *such that (i)* $\mu_g \in (c_1, 1-c_1)$ *for all* $g \in \{r,b\}$*, and (ii)* $\operatorname{ess\,sup}_{x\in\mathcal{X}} \mathbb{E}^*\left[(L_d^{g,\iota})^2|X=x\right] \leq c_2$ *for all* $d \in \{0,1\}$*,* $g \in \{r,b\}$*, and* $\iota \in \{A,F\}$.

DEFINITION 2—Algorithm space: An *algorithm* is a Borel-measurable map $a : \mathcal{X} \to [0,1]$, where $a(x)$ represents the probability of choosing $D=1$ given $X=x$. Denote $\mathcal{A}(\mathcal{X})$ the set of all such measurable maps. Write $\mathcal{A}$ when the covariate space is clear from context.

For each $a \in \mathcal{A}$ and $g \in \{r,b\}$, define the group-wise expected loss associated with loss function $\ell^\iota, \iota \in \{A,F\}$, as:

$$e_g^\iota(a) \equiv \mathbb{E}^*[\ell^\iota(D,Y^*)|G=g] = \mathbb{E}^*[a(X)\theta_1^{g,\iota}(X) + (1-a(X))\theta_0^{g,\iota}(X)]. \tag{3}$$

LLMO (Appendix O.1) derive the FA-frontier and study its properties when the loss function used to measure accuracy differs from that used to measure fairness, which is the framework used in this paper.[4] Here we adapt their definitions to our notation. They work in $\mathbb{R}^2$ with a feasible set defined with respect to the *accuracy* loss:

$$\mathcal{E}^A = \{(e_r^A(a), e_b^A(a)) : a \in \mathcal{A}\} \subset \mathbb{R}^2. \tag{4}$$

LLMO measure unfairness of algorithm a through $|f(a)|$, with $f : \mathcal{A} \to \mathbb{R}$ a linear functional. Their framework includes as a special case $f(a) = e_r^F(a) - e_b^F(a)$, the disparity in

[4] They use $\ell : \{0,1\} \times \mathcal{Y} \to \mathbb{R}$ to measure accuracy (what we call ℓ^A), and $\tilde{\ell} : \{0,1\} \times \mathcal{Y} \to \mathbb{R}$ to measure fairness (what we call ℓ^F).

expected fairness loss between groups, and for simplicity we specialize our discussion to that case. For each pair of expected accuracy losses $c = (c_r, c_b) \in \mathcal{E}^A$, define the minimal achievable unfairness as

$$d(c) \equiv \min\{|f(a)| : (e_r^A(a), e_b^A(a)) = c,\ a \in \mathcal{A}\}. \tag{5}$$

DEFINITION 3—LLMO's FA-dominance and FA-frontier : For $c, c' \in \mathcal{E}^A$, we say $c' \succ_{FA^{\text{LLMO}}} c$ if $c'_r \leq c_r,\ c'_b \leq c_b,\ d(c') \leq d(c)$, with at least one inequality strict. Then the FA-frontier relative to the accuracy loss is:

$$\mathcal{F}^A \equiv \{c \in \mathcal{E}^A : \nexists c' \in \mathcal{E}^A \text{ such that } c' \succ_{FA^{\text{LLMO}}} c\}. \tag{6}$$

Given the group-wise expected losses in (3), induced by algorithm $a \in \mathcal{A}$ for $\iota \in \{A, F\}$, we propose to work with the *feasible set* for $\left\{\left(e_r^A(a), e_b^A(a), e_r^F(a) - e_b^F(a)\right) : a \in \mathcal{A}\right\}$, where we explicitly include a coordinate measuring the difference in expected fairness loss for the two groups. Doing so allows us to characterize the properties of the FA frontier, extending LLMO's results, and to provide a representation of it that is amenable to estimation and inference, even in the presence of selectively observed data.

DEFINITION 4—Feasible set: Given $\mathbb{P}^*$ and $\mathcal{A}$, as a ranges in $\mathcal{A}$, the *feasible set* for $\left(e_r^A(a), e_b^A(a), e_r^F(a) - e_b^F(a)\right)$ is

$$\mathcal{E} \equiv \left\{(\varepsilon_1, \varepsilon_2, \varepsilon_3) \in \mathbb{R}^3 : \exists a \in \mathcal{A} \text{ such that } \varepsilon_1 = e_r^A(a), \varepsilon_2 = e_b^A(a), \varepsilon_3 = e_r^F(a) - e_b^F(a)\right\}.$$

To simplify notation, we omit the dependence of $\mathcal{E}$ on $(\mathbb{P}^*, \mathcal{A})$. In Lemma B.2, we show that under Assumptions 1-2, $\mathcal{E}$ is non-empty, compact, and convex. Definition 4 leads to an alternative formulation of FA-dominance and to an FA-frontier in $\mathbb{R}^3$, as follows:[5]

DEFINITION 5—FA-dominance and FA-frontier: Given $\varepsilon, \varepsilon' \in \mathcal{E}$, $\varepsilon' \succ_{FA} \varepsilon$ if $\varepsilon'_1 \leq \varepsilon_1,\ \varepsilon'_2 \leq \varepsilon_2,\ |\varepsilon'_3| \leq |\varepsilon_3|$, with at least one inequality strict. The FA-frontier is:

$$\mathcal{F} \equiv \{\varepsilon \in \mathcal{E} : \nexists\, \varepsilon' \in \mathcal{E} \text{ such that } \varepsilon' \succ_{FA} \varepsilon\}. \tag{7}$$

[5] Auerbach et al. (2024, p.8) use a similar definition of dominance.

In the next section we show that, in fact, our FA-frontier and LLMO's FA-frontier coincide in what we argue is the only way that matters.

2.2. *Support Function-Based Characterization of the FA-Frontier*

To begin, we characterize (similarly to Liu and Molinari, 2026, Propositions 3.1-3.2) the support function of $\mathcal{E}$, denoted $h_{\mathcal{E}}(q)$ for q a direction vector in $\mathbb{S}^2 \equiv \{q \in \mathbb{R}^3 : \|q\| = 1\}$. This functional turns out to be crucial for all our results. We assume:

ASSUMPTION 2—Margin condition: *There exist constants $m \in (0, 1]$ and $C < \infty$ such that for every $\delta > 0$, $\sup_{q \in \mathbb{S}^2} \mathbb{P}^*\big(|q^\intercal \Delta\boldsymbol{\theta}(X)| \leq \delta\big) \leq C\delta^m$.*

THEOREM 2.1—Support function of $\mathcal{E}$: *Under Assumptions 1-2, the support function of $\mathcal{E}$, $h_{\mathcal{E}}(q)$, and its gradient with respect to $q \in \mathbb{S}^2$, are*

$$h_{\mathcal{E}}(q) \equiv \sup_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon = \mathbb{E}^*\big[q^\intercal \mathbf{L}_0 + q^\intercal(\mathbf{L}_1 - \mathbf{L}_0)\mathbb{1}\{q^\intercal \Delta\boldsymbol{\theta}(X) > 0\}\big], \tag{8}$$

$$\nabla_q h_{\mathcal{E}}(q) = \mathbb{E}^*\big[\mathbf{L}_0 + (\mathbf{L}_1 - \mathbf{L}_0)\mathbb{1}\{q^\intercal \Delta\boldsymbol{\theta}(X) > 0\}\big]. \tag{9}$$

As $\nabla_q h_{\mathcal{E}}(q)$ is a singleton for all $q \in \mathbb{S}^2$, the set $\mathcal{E}$ is strictly convex.

As the goal of this section is to characterize $\mathcal{F}$, given $\varepsilon^* = (\varepsilon_1^*, \varepsilon_2^*, \varepsilon_3^*) \in \mathcal{E}$, we define

$$\mathcal{C}(\varepsilon^*) \equiv \{\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3) \in \mathbb{R}^3 : \varepsilon_1 \leq \varepsilon_1^*,\ \varepsilon_2 \leq \varepsilon_2^*,\ |\varepsilon_3| \leq |\varepsilon_3^*|\}, \tag{10}$$

the set of group-wise expected accuracy losses and fairness disparity $\varepsilon \in \mathbb{R}^3$—whether or not they are feasible—that are both weakly more accurate and weakly fairer than ε^*. In Lemma B.1, we show that the support function of $\mathcal{C}(\varepsilon^*)$ equals:

$$h_{\mathcal{C}(\varepsilon^*)}(q) \equiv \sup_{\varepsilon \in \mathcal{C}(\varepsilon^*)} q^\intercal \varepsilon = \begin{cases} q_1\varepsilon_1^* + q_2\varepsilon_2^* + |q_3||\varepsilon_3^*| & \text{if } q_1 \geq 0,\ q_2 \geq 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{11}$$

Given an algorithm $a^* \in \mathcal{A}$, we now show that $(e_r^A(a^*), e_b^A(a^*), e_r^F(a^*) - e_b^F(a^*)) \in \mathcal{F}$ if and only if $(e_r^A, e_b^A) \in \mathcal{F}^A$ with $e_r^A = e_r^A(a^*)$ and $e_b^A = e_b^A(a^*)$, and $|e_r^F(a^*) - e_b^F(a^*)|$

equals the minimum absolute disparity in (5) associated with (e_r^A, e_b^A). Define the projection $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ by $\pi(x_1, x_2, x_3) = (x_1, x_2)$, and let $\pi(\mathcal{E}) = \{\pi(\varepsilon) : \varepsilon \in \mathcal{E}\}$. Then it is immediate from the definitions that $\pi(\mathcal{E}) = \mathcal{E}^A$.

THEOREM 2.2: *Let* $\varepsilon^* = (\varepsilon_1^*, \varepsilon_2^*, \varepsilon_3^*) \in \mathcal{E}$ *and define* $\widetilde{\mathbb{S}}^2 \equiv \{q \in \mathbb{S}^2 : q_1 \geq 0,\ q_2 \geq 0\}$. *Under Assumptions 1-2, the following are equivalent:*

(i) $\varepsilon^* \in \mathcal{F}$;

(ii) $\mathcal{E} \cap \mathcal{C}(\varepsilon^*) = \{\varepsilon^*\}$;

(iii) $\min_{q \in \widetilde{\mathbb{S}}^2} \left(h_{\mathcal{C}(\varepsilon^*)}(q) + h_{\mathcal{E}}(-q)\right) = 0$;

(iv) $\pi(\varepsilon^*) = (\varepsilon_1^*, \varepsilon_2^*) \in \mathcal{F}^A$ *and* $|\varepsilon_3^*| = d((\varepsilon_1^*, \varepsilon_2^*))$.

An immediate consequence of Theorem 2.2 is that $\mathcal{F}$ captures as many features of the trade-off between accuracy and fairness as $\mathcal{F}^A$ does in LLMO's original contribution, while preserving convexity of the feasible set and yielding a simple characterization of the FA-frontier.[6] Indeed, using the equivalence $(i) \Leftrightarrow (iii)$ in Theorem 2.2, we obtain

$$\mathcal{F} = \left\{ \varepsilon^* \in \mathcal{E} : \min_{q \in \widetilde{\mathbb{S}}^2} \left[h_{\mathcal{C}(\varepsilon^*)}(q) + h_{\mathcal{E}}(-q)\right] = 0 \right\}. \tag{12}$$

When the same loss function is used to measure accuracy and fairness, as in LLMO's main text and in Liu and Molinari (2026), it is simple to characterize when the FA-frontier coincides with the Pareto frontier.[7] In our framework, the latter is given by:

$$\mathcal{PF} \equiv \left\{ \varepsilon \in \mathcal{E} : \nexists \varepsilon' \in \mathcal{E} \text{ s.t. } \varepsilon_1' \leq \varepsilon_1,\ \varepsilon_2' \leq \varepsilon_2, \text{ with at least one strict inequality} \right\}. \tag{13}$$

Here we provide a simple condition based on the support function, under which $\mathcal{PF} = \mathcal{F}$. The characterization is valuable because when the FA-frontier does not coincide with the Pareto frontier, the trade-off between fairness and accuracy becomes substantial: a policy

[6] In fact, $\mathcal{F}$ yields a finer characterization. Consider algorithms $a, a' \in \mathcal{A}$ such that $\varepsilon := \varepsilon(a) = (0.5, 0.5, 1)$ and $\varepsilon' := \varepsilon(a') = (0.5, 0.5, 0)$. Then $\varepsilon' \succ_{FA} \varepsilon$ but $c' \not\succ_{FA^{\text{LLMO}}} c$ for $c = (\varepsilon_1, \varepsilon_2)$ and $c' = (\varepsilon_1', \varepsilon_2')$.

[7] In that context, $\ell^A = \ell^F$ and $\varepsilon_g^A = \varepsilon_g^F$ for $g \in \{r, b\}$, so the feasible set $\mathcal{E}$ can be defined in $\mathbb{R}^2$. Denoting R the vector in $\mathcal{E}$ that minimizes e_r and B that minimizing e_b, LLMO show in this case that the Pareto and the FA frontiers coincide if and only if R and B lie on opposite sides of the 45^o line.

maker who values fairness may choose to deploy an algorithm that is Pareto dominated in accuracy loss, to ameliorate its performance relative to fairness loss.

THEOREM 2.3: *Under Assumptions 1-2, $\mathcal{PF} = \mathcal{F}$ if and only if for all $\varepsilon^* \in \mathcal{F}$ there exists $\alpha \in [0,1]$ such that for $q(\alpha) = \frac{(\alpha, 1-\alpha, 0)}{\|(\alpha, 1-\alpha, 0)\|}$, $\varepsilon^* = \arg\min_{\eta \in \mathcal{E}} q(\alpha)^\intercal \eta$.*

In Appendix B.1.1, we show that $\mathcal{PF} \subsetneq \mathcal{F}$ if $\mathcal{E}$ has no kinks and lies completely to one side of the hyperplane $\{\varepsilon_3 = 0\}$ (which corresponds to the hyperplane where $e_r^F = e_b^F$), see Theorem B.1 and Corollary B.1; and that $\mathcal{PF} \subsetneq \mathcal{F}$ if $\mathcal{E}$ crosses the hyperplane $\{\varepsilon_3 = 0\}$ but $\mathcal{PF}$ lies completely to one side of it, see Theorem B.2. Hence, the support function characterization in Theorem 2.3 is particularly useful in the case where $\mathcal{E}$ and $\mathcal{PF}$ both cross the hyperplane $\{\varepsilon_3 = 0\}$.

3. SELECTIVELY OBSERVED LABELS

Having fully characterized the FA-frontier in the idealized case that $(Y^*, X, G, Z) \sim \mathbb{P}^*$ is observed, we turn to our goal of providing identification results and inference procedures for $\mathcal{F}$ when labels are selectively observed: the outcome we observe is $Y = ZY^*$, for Z a selection indicator. This problem has been studied by Lakkaraju et al. (2017), Kleinberg et al. (2017), Rambachan et al. (2025), Khan et al. (2025), among others. A classical motivation arises when evaluating whether algorithmic policies can improve upon human decision-making, in which case selective labels are compounded by the possibility that Z is determined by a decision maker (e.g., a judge) who observes information beyond the recorded covariates made available to the algorithm. In this case, for a specific choice of loss functions, we report partial identification results. We then provide point identification results and statistical inference procedures for the case where label observability is induced by deployment of a decision rule in $\mathcal{A}(\mathcal{X})$, so that Z is determined by an algorithm that uses only the inputs in $\mathcal{X}$.

3.1. *Partial Identification of the FA-Frontier*

In this section, we allow Z to depend on variables unavailable to the algorithms in $\mathcal{A}(\mathcal{X})$. This creates substantial challenges for identification and inference, because even for a fixed

algorithm $a^* \in \mathcal{A}(\mathcal{X})$, $\varepsilon^* := (\varepsilon_r^A(a^*), \varepsilon_b^A(a^*), \varepsilon_r^F(a^*) - \varepsilon_b^F(a^*))$ is only partially identified. Consequently, extra care needs to be put in making sure, when characterizing what can be learned about $\mathcal{F}$ in (12), to correctly account for partial identification of both $\mathcal{E}$ and ε^*.

To make progress on this task, we specialize our analysis to the case where Y^* is a binary scalar variable, the accuracy loss measures classification error, $\ell^A(d, y) = \mathbb{1}\{d \neq y\}$, and the fairness loss measures statistical parity, $\ell^F(d, y) = \mathbb{1}\{d = 1\}$. Write $Y = ZY^*$ for the observed binary outcome. In this case the fairness coordinate depends only on the decision rule and the observable distribution of (X, G), so lack of point identification comes entirely from the selectively observed outcomes entering the two accuracy coordinates.

Denote the unobserved conditional expectation of Y^* when $Z = 0$ by

$$\lambda_g^*(x) \equiv \mathbb{E}^*[Y^*|X = x, G = g, Z = 0], \qquad g \in \{r, b\}$$

and $\lambda^*(x) = (\lambda_r^*(x), \lambda_b^*(x))^\intercal$. Without restrictions on the selection process, λ^* is not point identified. Because $Y^* \in \{0, 1\}$, every measurable map $\lambda = (\lambda_r, \lambda_b) : \mathcal{X} \to [0, 1]^2$ is compatible with $\mathbb{P}$, the observed law of (Y, X, G, Z).[8] Let $\Lambda \equiv \left\{\lambda = (\lambda_r, \lambda_b) : \mathcal{X} \to [0, 1]^2\right\}$ denote this class of functions.

We next derive the sharp identification region for $h_{\mathcal{E}}(q)$ in direction $q \in \mathbb{S}^2$. To do so, we need to introduce some notation. Define the observable functions

$$A_0(x) \equiv \mathbb{E}\left[\left(\frac{ZY\mathbb{1}\{G=r\}}{\mu_r}, \frac{ZY\mathbb{1}\{G=b\}}{\mu_b}, 0\right)^\intercal \middle| X = x\right], \tag{14}$$

$$A_1(x) \equiv \mathbb{E}\left[\left(\frac{(1-ZY)\mathbb{1}\{G=r\}}{\mu_r}, \frac{(1-ZY)\mathbb{1}\{G=b\}}{\mu_b}, \frac{\mathbb{1}\{G=r\}}{\mu_r} - \frac{\mathbb{1}\{G=b\}}{\mu_b}\right)^\intercal \middle| X = x\right], \tag{15}$$

$$B(x) \equiv \mathbb{E}\left[\begin{pmatrix} \frac{(1-Z)\mathbb{1}\{G=r\}}{\mu_r} & 0 \\ 0 & \frac{(1-Z)\mathbb{1}\{G=b\}}{\mu_b} \\ 0 & 0 \end{pmatrix} \middle| X = x\right]. \tag{16}$$

Lemma A.1 shows that the functions $\boldsymbol{\theta}_d(\cdot)$ in Definition 1 associated with distribution $\lambda \in \Lambda$ can be expressed as $\boldsymbol{\theta}_0(x; \lambda) = A_0(x) + B(x)\lambda(x)$ and $\boldsymbol{\theta}_1(x; \lambda) = A_1(x) - B(x)\lambda(x)$.

[8] By compatible with the observed data we mean that it yields a law of (Y^*, X, G, Z) that reproduces the observed distribution of (Y, X, G, Z) and satisfies $Y = ZY^*$ almost surely.

Hence, the support function of the feasible set associated with distribution λ is

$$h_{\mathcal{E}}(q;\lambda) = \mathbb{E}\left[\max\left\{q^{\intercal}\left(A_0(X) + B(X)\lambda(X)\right), q^{\intercal}\left(A_1(X) - B(X)\lambda(X)\right)\right\}\right]. \tag{17}$$

Accordingly, the sharp identified set for $h_{\mathcal{E}}(q)$ in direction q is $\{h_{\mathcal{E}}(q;\lambda) : \lambda \in \Lambda\}$.

For a fixed algorithm a^*, let $A_{d,j}(x)$ denote the jth coordinate of $A_d(x)$, $d \in \{0,1\}$, in (14)-(15) and define

$$C(q;a^*) \equiv \sum_{j=1}^{2} q_j \mathbb{E}\left[a^*(X)A_{1,j}(X) + (1-a^*(X))A_{0,j}(X)\right] + |q_3|\,\left|\mathbb{E}\left[a^*(X)A_{1,3}(X)\right]\right|. \tag{18}$$

Then we can express $h_{\mathcal{C}(\varepsilon^*)}$ in (11) as

$$h_{\mathcal{C}(\varepsilon^*)}(q) = C(q;a^*) + \mathbb{E}^*\left[(1-2a^*(X))\, q^{\intercal}B(X)\lambda^*(X)\right] \tag{19}$$

Theorem 2.2 continues to characterize frontier membership for given λ, where crucially we use the same candidate unobserved distribution λ both in obtaining $h_{\mathcal{E}}(\cdot)$ and $h_{\mathcal{C}(\varepsilon^*)}(\cdot)$. In particular, if $\varepsilon^*(a^*;\lambda)$ denotes the vector of expected losses induced by a^* under unobserved distribution λ, then a^* lies on the FA-frontier under λ if and only if

$$\min_{q\in\tilde{\mathbb{S}}^2}\left[h_{\mathcal{C}(\varepsilon^*(a^*;\lambda))}(q) + h_{\mathcal{E}}(-q;\lambda)\right] = 0.$$

The next result shows how to check whether *some* distribution for the unobserved labels places a^* on the frontier through an optimization problem that involves functionals of the observed distribution $\mathbb{P}$.

THEOREM 3.1: *Let Assumptions 1–2 hold. Suppose* $Y^* \in \{0,1\}$, $\ell^A(d,y) = \mathbb{1}\{d \neq y\}$, *and* $\ell^F(d,y) = \mathbb{1}\{d = 1\}$. *Then a given algorithm* a^* *lies on the FA-frontier for some distribution consistent with the observed data if and only if*

$$\min_{q\in\tilde{\mathbb{S}}^2} \mathbb{E}\left[\max\{J_0(\lambda_1;q,a^*), (1-a^*)J_0(\lambda_0;q,a^*) + a^*J_1(\lambda_0;q,a^*), J_1(\lambda_0;q,a^*)\}\right] = 0. \tag{20}$$

where $\lambda_0 = (0,0)^{\intercal}$, $\lambda_1 = (1,1)^{\intercal}$, *and*

$$J_0(\lambda;q,a^*) \equiv C(q;a^*) - q^{\intercal}A_0(X) - 2a^*(X)q^{\intercal}B(X)\lambda,$$

$$J_1(\lambda; q, a^*) \equiv C(q; a^*) - q^\intercal A_1(X) + 2(1 - a^*(X))q^\intercal B(X)\lambda. \tag{21}$$

Theorem 3.1 shows that once the source of partial identification is reduced to the functional λ, the search over the infinite-dimensional class Λ collapses to a finite-dimensional criterion involving only the corner distributions λ_0 and λ_1.

REMARK 3.1: In work in progress, we obtain a debiased machine learning estimator for $J_d(\lambda; q, a^*)$, $d \in \{0, 1\}$, develop an inference procedure to test the hypothesis that a given algorithm is on the FA-frontier, and explore extending Theorem 3.1 to other loss functions.

3.2. *Point Identification Under Missing at Random Assumptions*

We represent the observed testing data by the random vector (Y, G, X, Z) with distribution denoted $\mathbb{P}$. While $\mathbb{P}$ is point identified by the observed data, in the absence of additional assumptions, the distribution $\mathbb{P}^*$ of (Y^*, G, X, Z) is not. Consequently, the support function $h_{\mathcal{E}}(q)$ is not point identified either, as the ideal labels and their conditional expectations ($\mathbf{L}_d$ and $\boldsymbol{\theta}_d(X)$, respectively) are a function of Y^* and $\mathbb{P}^*$.

We leverage the fact that all algorithms in $\mathcal{A}(\mathcal{X})$ can only use the covariates X for training (and no other unobserved variables) to argue plausibility of a selection on observables assumption that we use to point identify the distribution $\mathbb{P}^*$.

ASSUMPTION 3—Ideal labels missing at random (MAR): $(Y^*, G) \perp Z|X$ *and* $\boldsymbol{\pi}(X) = \mathbb{E}[Z|X] \in (0, 1)$ *a.s.*

Assumption 3 may look stronger than the typical missing at random (unconfoundedness or selection on observables) assumption, which might instead require $Y^* \perp Z|X, G$. In work in progress, we establish point identification of the support function of $\mathcal{E}$ and derive a debiased machine learning estimator for it under this weaker assumption, along with its asymptotic distribution. Here we work with Assumption 3 because it yields lighter notation and simpler expressions. We argue that, in many settings of interest, the additional requirement $G \perp Z|X$ is a natural consequence of how Z is generated. Specifically, suppose that $Z = D$, where $D \in \{0, 1\}$ is the binary decision an individual receives

(e.g., granted or denied a loan; granted or denied bail), and that this decision is produced by a deployed decision rule in $\mathcal{A}(\mathcal{X})$ that takes X as input and, following regulations, does not use G or apply any group-dependent post-processing. Then, for all $g \in \{r, b\}$, $\mathbb{P}(D = 1|X, G = g) = \mathbb{P}(D = 1|X) = a(X)$, so that $G \perp Z|X$ holds. Combined with the standard selection on observables condition, $Y^* \perp Z|X, G$, this yields Assumption 3. The more restrictive content of Assumption 3 is therefore the requirement $Y^* \perp Z|X, G$. This condition is plausible when the assignment mechanism that determines Z depends on recorded variables capturing the main drivers of both selection and outcomes, and when discretion at the decision-making stage is limited or can be accounted for. In such settings, conditioning on (X, G) may reasonably approximate conditioning on the information that drives selection, so that Z does not convey additional information about Y^* beyond (X, G). At the same time, $Y^* \perp Z|X, G$ may fail when the assignment process exploits unrecorded "soft information" (e.g., interviews, free text, documents, or unlogged history) or when discretion is systematically related to expected outcomes. In these cases, Z may remain informative about Y^* even after conditioning on (X, G), and point identification need not hold, motivating our partial identification analysis-in-progress as per Remark **??**.

The ideal labels $\mathbf{L}_d$ are a non-stochastic function of Y^* and G, and $Z\mathbf{L}_d$ is observable. With that, under Assumption 3 the support function of $\mathcal{E}$ is point identified through a standard inverse propensity weighting identity, the proof of which we provide in the Appendix for completeness.

LEMMA 3.1: *Let Assumptions 1 and 3 hold. Then for* $\Delta\mathbf{L} \equiv \mathbf{L}_1 - \mathbf{L}_0$,

$$\boldsymbol{\theta}_d(X) = \mathbb{E}\left[\frac{Z\mathbf{L}_d}{\boldsymbol{\pi}(X)}\middle| X\right], \tag{22}$$

$$h_{\mathcal{E}}(q) = \mathbb{E}\left[\frac{q^\intercal \mathbf{L}_0 Z}{\boldsymbol{\pi}(X)} + \left(\frac{q^\intercal \Delta\mathbf{L} Z}{\boldsymbol{\pi}(X)}\right) \mathbb{1}\{q^\intercal \Delta\boldsymbol{\theta}(X) > 0\}\right], \quad \textit{for all } q \in \mathbb{S}^2, \tag{23}$$

$$e^\iota_g(a) = \mathbb{E}\left[\frac{a(X)L_1^{g,\iota} Z}{\boldsymbol{\pi}(X)} + \frac{(1-a(X))L_0^{g,\iota} Z}{\boldsymbol{\pi}(X)}\right], \quad g = r, b \textit{ and } \iota = A, F. \tag{24}$$

3.3. *A Debiased Machine Learning Estimator*

Our goal is to estimate the support function $h_{\mathcal{E}}(q)$ using a random sample $\{(Y_i, G_i, X_i, Z_i) : 1 \leq i \leq n\}$ drawn from the distribution $\mathbb{P}$. By Lemma 3.1, Eqs. (22)-(23), we can write

$$h_{\mathcal{E}}(q) = \mathbb{E}\left[\frac{q^{\intercal}\mathbf{L}_{0,i}Z_i}{\boldsymbol{\pi}(X_i)} + \left(\frac{q^{\intercal}\Delta\mathbf{L}_i Z_i}{\boldsymbol{\pi}(X_i)}\right)\mathbb{1}\{q^{\intercal}\Delta\boldsymbol{\theta}(X_i) > 0\}\right],$$

where $\boldsymbol{\pi}(X)$ and $\Delta\boldsymbol{\theta}(X)$ are nuisance functions that need to be estimated in a first-step. As the ideal labels $\mathbf{L}$ depend on $\boldsymbol{\mu} \equiv (\mu_r, \mu_g)$ (see Eqs. 1-2), in the first step we also need to estimate $\boldsymbol{\mu}$. In order to put forward an estimator for $h_{\mathcal{E}}(q),\ q \in \mathbb{S}^2$, that is locally robust to the fist-step estimation error, we rely on the debiased machine learning (DML) approach. Consider the estimand

$$h_{\mathcal{E}}(q) = \mathbb{E}\left[\zeta_i(q; \mathbf{L}, \boldsymbol{\eta})\right], \tag{25}$$

where $\boldsymbol{\eta} = (\Delta\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\theta}_0)$ is a vector of nuisance functions and

$$\zeta_i(q; \mathbf{L}, \boldsymbol{\eta}) \equiv \frac{q^{\intercal}\mathbf{L}_{0,i}Z_i}{\boldsymbol{\pi}(X_i)} + \left(\frac{q^{\intercal}\Delta\mathbf{L}_i Z_i}{\boldsymbol{\pi}(X_i)}\right)\mathbb{1}\{q^{\intercal}\Delta\boldsymbol{\theta}(X_i) > 0\} + \boldsymbol{\alpha}^h(q; X_i)\left(1 - \frac{Z_i}{\boldsymbol{\pi}(X_i)}\right), \tag{26}$$

$$\boldsymbol{\alpha}^h(q; X) = q^{\intercal}\boldsymbol{\theta}_0(X) + q^{\intercal}\Delta\boldsymbol{\theta}(X)\mathbb{1}\{q^{\intercal}\Delta\boldsymbol{\theta}(X) > 0\}. \tag{27}$$

The local effect of the first-step function $\boldsymbol{\pi}$ on the original estimand $h_{\mathcal{E}}(q)$ in Eq. (25) is accounted for by $\boldsymbol{\alpha}^h(q; X_i)\left(1 - \frac{Z_i}{\boldsymbol{\pi}(X_i)}\right)$. As we show in the proof of Theorem 3.2, under Assumptions 4 and 5 (presented below) it is not necessary to account for the local effect of the first-step function $\Delta\boldsymbol{\theta}$ on $h_{\mathcal{E}}(q)$. Finally, we note that the estimand in Eq. (25) is based on $\mathbf{L}$, which by Eq. (1) relies on the population probabilities $\mu_g, g \in \{r, b\}$.

Following the DML approach (e.g., Chernozhukov et al., 2018), we first randomly split the indices $[n] = \{1, \ldots, n\}$ into K equal-sized folds $\mathcal{I}_k$, i.e., $\cup_{k=1}^{K}\mathcal{I}_k = [n]$. We denote by n_k the size of $\mathcal{I}_k$.[9] We then estimate $\widehat{\boldsymbol{\eta}}(X_i)$ as $\widehat{\boldsymbol{\eta}}_k(X_i)$ for $i \in \mathcal{I}_k$ and $k = 1, \ldots, K$, where $\widehat{\boldsymbol{\eta}}_k(\cdot)$ is estimated using all data except the portion with indices in fold $\mathcal{I}_k$. Finally, the preliminary estimates of the nuisance functions are used to calculate an estimator for $h_{\mathcal{E}}(q)$

[9] When n is not divisible by K, the number of observations in some folds will be $\lfloor n/K \rfloor$ while in others $\lfloor n/K \rfloor + 1$, where $\lfloor n/K \rfloor$ is the greatest integer less than or equal to n/K.

based on the estimand defined in Eq. (25) as follows

$$\widehat{h}_{\mathcal{E}}(q;\widehat{\mathbf{L}},\widehat{\boldsymbol{\eta}}) \equiv \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i\in\mathcal{I}_k}\zeta_i(q;\widehat{\mathbf{L}},\widehat{\boldsymbol{\eta}}_k)\right) = \frac{1}{n}\sum_{i=1}^{n}\zeta_i(q;\widehat{\mathbf{L}},\widehat{\boldsymbol{\eta}}), \tag{28}$$

where $\widehat{\mathbf{L}}$ is defined as in Eqs. (1)-(2) with μ_g replaced by $\widehat{\mu}_g = n^{-1}\sum_{i=1}^{n}\mathbb{1}\{G_i = g\}$.

We also propose a DML estimator for $\varepsilon^* = \varepsilon(a^*) \in \mathcal{E}$. By Eq. (24) in Lemma 3.1 and Definition 3, we can write

$$\varepsilon^* = \mathbb{E}[\xi_i(a^*;\mathbf{L},\boldsymbol{\eta})],$$

where, for a given algorithm $a \in \mathcal{A}$,

$$\xi_i(a;\mathbf{L},\boldsymbol{\eta}) \equiv (\mathbf{L}_{0,i} + a(X_i)(\mathbf{L}_{1,i} - \mathbf{L}_{0,i}))\,\frac{Z_i}{\boldsymbol{\pi}(X_i)} + \boldsymbol{\alpha}^e(a;X_i)\left(1 - \frac{Z_i}{\boldsymbol{\pi}(X_i)}\right), \tag{29}$$

$$\boldsymbol{\alpha}^e(a;X) \equiv \boldsymbol{\theta}_0(X) + a(X)\Delta\boldsymbol{\theta}(X), \tag{30}$$

and again the local effect of the first-step function $\boldsymbol{\pi}$ on the original estimand ε^* is accounted for by $\boldsymbol{\alpha}^e(X_i)\big(1 - \frac{Z_i}{\boldsymbol{\pi}(X_i)}\big)$. Using this notation, we propose to estimate ε^* by

$$\widehat{\varepsilon}^* \equiv \frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{n_k}\sum_{i\in\mathcal{I}_k}\xi_i(a^*;\widehat{\mathbf{L}},\widehat{\boldsymbol{\eta}}_k)\right) = \frac{1}{n}\sum_{i=1}^{n}\xi_i(a^*;\widehat{\mathbf{L}},\widehat{\boldsymbol{\eta}})\ , \tag{31}$$

where $\widehat{\mathbf{L}}$ are the feasible labels and $\widehat{\boldsymbol{\eta}}_k$'s are estimates of the nuisance function using cross-fitting, as we described above.

REMARK 3.2: DML relies on two elements: a moment function that satisfies a Neyman orthogonality condition and a nuisance function estimator based on cross-fitting that is sufficiently good (its convergence rate is faster than $o_p(n^{-1/4})$). With these two elements in place, the first-step estimation error does not affect the asymptotic distribution of the DML estimator, as we show in the proof of Theorem 3.2. In other words, the DML estimator is as good as an oracle estimator that knows the nuisance functions.

3.4. *Asymptotic Distribution of the DML Estimator*

For the estimation error of the nuisance functions estimated in first step not to affect the asymptotic distribution of the DML estimator, one needs such estimators to be sufficiently well-behaved. To ensure this, we ask that $\Delta\boldsymbol{\theta}(X)$ and $\boldsymbol{\pi}(X)$ satisfy an assumption similar to but weaker than Liu and Molinari (2026, Assumption 3).

ASSUMPTION 4: *There is a known partition of* X,

$$X = (X_1, X_2),$$

with $X_1 \in \mathcal{X}_1 \subset \mathbb{R}^p$ *and* $X_2 \in \mathcal{X}_2 \subset \mathbb{R}^{d_X - p}$ *such that:*

(a) *The density of* $|q^\intercal \Delta\boldsymbol{\theta}(X)|$ *conditional on* X_2 *is uniformly bounded in* $q \in \mathbb{S}^2$, *that is,*

$$\mathbb{P}\left(|q^\intercal \Delta\boldsymbol{\theta}(X)| \leq \delta | X_2\right) \leq C_\delta \delta$$

for a given positive constant C_δ.

(b) *There is a constant* $c > 0$ *such that*

$$\inf_{q \in \mathbb{S}^2} \mathbb{V}ar[|q^\intercal \Delta\boldsymbol{\theta}(X)||X_2] \geq c\,.$$

(c) *There are unknown functions* $\nu : \mathcal{X}_1 \to \mathbb{R}^{d_\nu}$ *and* $\gamma : \mathcal{X}_2 \to \mathbb{R}^{d_\gamma}$, *and a known function* $F : \mathbb{R}^{d_\nu} \times \mathbb{R}^{d_\gamma} \times \mathbb{R}^2 \to [\epsilon, 1-\epsilon] \times \mathbb{R}^3$, *such that:*

$$(\boldsymbol{\pi}(X),\ \Delta\boldsymbol{\theta}(X)) = F(\nu(X_1), \gamma(X_2), \mu)$$

and the partial derivatives of F *are uniformly bounded for any* $\nu \in \mathbb{R}^{d_\nu}$, $\gamma \in \mathbb{R}^{d_\gamma}$, *and* $\mu \in \mathbb{B}_\epsilon(\mu_r, \mu_b) = \{\mu \in \mathbb{R}^2 : ||\mu - (\mu_r, \mu_b)|| < \epsilon\}$, *that is*

$$\sup_{\nu \in \mathbb{R}^{d_\nu}} \sup_{\gamma \in \mathbb{R}^{d_\gamma}} \sup_{\mu \in \mathbb{B}_\epsilon(\mu_r, \mu_b)} ||DF(\nu, \gamma, \mu)|| < \infty\,.$$

Let $\mathcal{F}_k = \sigma((Y_i, G_i, X_i, Z_i) : i \notin \mathcal{I}_k)$ be the σ-algebra generated by the data used for the estimation of the nuisance function $\boldsymbol{\eta} = (\Delta\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\theta}_0)$. Let $\widehat{\boldsymbol{\mu}}_k$ be an estimator for $(\mu_r, \mu_b)^\intercal$

that uses only the data outside fold $\mathcal{I}_k$ (like we do for the estimator $\widehat{\boldsymbol{\eta}}_k$):

$$\widehat{\boldsymbol{\mu}}_k = \left(\tfrac{1}{n-n_k} \sum_{i \notin \mathcal{I}_k} \mathbb{1}\{G_i = r\},\ \tfrac{1}{n-n_k} \sum_{i \notin \mathcal{I}_k} \mathbb{1}\{G_i = b\} \right)^\intercal .$$

We assume that the estimators for $\Delta\boldsymbol{\theta}(X)$ and $\boldsymbol{\pi}(X)$ share the same structure as in Assumption 4. Furthermore, we treat $\nu(X_1)$ as a low-dimensional nonparametric component that can be estimated uniformly well over $\mathcal{X}_1$, whereas $\gamma(X_2)$ captures a high-dimensional component that can be estimated sufficiently well in the sense that its mean squared error converges to zero sufficiently fast. We assume that $\boldsymbol{\theta}_0$ can be estimated sufficiently well.

ASSUMPTION 5: *There exist estimators $\widehat{\nu}_k$, $\widehat{\gamma}_k$, and $\widehat{\boldsymbol{\theta}}_{0,k}$ for $k = 1, \ldots, K$ such that*

(a) $(\widehat{\boldsymbol{\pi}}_k(X),\ \Delta\widehat{\boldsymbol{\theta}}_k(X)) = F(\widehat{\nu}_k(X_1), \widehat{\gamma}_k(X_2), \widehat{\boldsymbol{\mu}}_k)$.

(b) $\sup_{x_1 \in \mathcal{X}_1} ||\widehat{\nu}_k(x_1) - \nu(x_1)|| = o_p(n^{-1/4})$.

(c) $\mathbb{E}\left[||\widehat{\gamma}_k(X_2) - \gamma(X_2)||^2 | \mathcal{F}_k\right] = o_p(n^{-1/2})$.

(d) $\mathbb{E}\left[||\widehat{\boldsymbol{\theta}}_{0,k}(X) - \boldsymbol{\theta}_0(X)||^2 | \mathcal{F}_k\right] = o_p(n^{-1/2})$.

Similarly to Liu and Molinari (2026, Theorem 4.1), our main asymptotic result shows that $\widehat{h}_{\mathcal{E}}(q; \widehat{\mathbf{L}}, \widehat{\boldsymbol{\eta}})$ converges to a Gaussian process uniformly in $q \in \mathbb{S}^2$, where the score $\zeta_i(q; \mathbf{L}, \boldsymbol{\eta})$ in Eq. (26) is the influence function that governs the part of the limit distribution of $\widehat{h}_{\mathcal{E}}(q; \widehat{\mathbf{L}}, \widehat{\boldsymbol{\eta}})$ due to $\widehat{\boldsymbol{\eta}}$, and the remaining part is attributed to estimating $\boldsymbol{\mu}$:

THEOREM 3.2: *Let Assumptions 1–5 hold. Then,*

$$\sqrt{n}\left(\widehat{h}_{\mathcal{E}}(q; \widehat{\mathbf{L}}, \widehat{\boldsymbol{\eta}}) - h_{\mathcal{E}}(q)\right) = \mathbb{G}[\zeta_i^*(q; \mathbf{L}, \boldsymbol{\eta})] + o_p(1) \quad \textit{in} \quad \ell^\infty(\mathbb{S}^2) \ ,$$

where

$$\zeta_i^*(q; \mathbf{L}, \boldsymbol{\eta}) = \zeta_i(q; \mathbf{L}, \boldsymbol{\eta}) + (\mathcal{M}_i q)^\intercal \nabla_q h_{\mathcal{E}}(q), \tag{32}$$

and $\mathcal{M}_i = \operatorname{diag}\left(\left[1 - \frac{\mathbb{1}\{G_i=r\}}{\mu_r}, 1 - \frac{\mathbb{1}\{G_i=b\}}{\mu_b}, \frac{\mathbb{1}\{G_i=b\}}{\mu_b} - \frac{\mathbb{1}\{G_i=r\}}{\mu_r}\right]\right)$.

REMARK 3.3: One could alternatively construct a DML estimator of $h_{\mathcal{E}}(q)$ using $\zeta_i^*(q; \boldsymbol{\eta}, \boldsymbol{\mu})$ instead of $\zeta_i(q; \boldsymbol{\eta}, \boldsymbol{\mu})$ as we do in Eq. (28). Doing so would be valid because

$\zeta_i^*(q;\boldsymbol{\eta},\boldsymbol{\mu})$ satisfies the Neyman orthogonality condition with respect to $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$. In practice, when $\widehat{\boldsymbol{\eta}}$ is obtained via cross-fitting and $\boldsymbol{\mu}$ (which enters $\mathbf{L}$ and $\mathcal{M}_i$) is estimated by plug-in, our estimator is numerically equivalent to that:

$$\widehat{h}_{\mathcal{E}}(q;\widehat{\mathbf{L}},\widehat{\boldsymbol{\eta}}) = n^{-1}\sum_{i=1}^{n}\zeta_i^*(q;\widehat{\mathbf{L}},\widehat{\boldsymbol{\eta}}).$$

Hence, our proposed estimator implicitly accounts for first-step estimation error in $\widehat{\boldsymbol{\mu}}$.

4. ASSESSING THE EXISTENCE OF A LESS DISCRIMINATORY ALTERNATIVE

When evaluating whether there exists a less discriminatory alternative (LDA) to a given algorithm, regulators and policymakers need to confront the fact that they can only rely on finite data to do so. More concretely, given an algorithm $a^* \in \mathcal{A}(\mathcal{X})$, we call another algorithm $\tilde{a} \in \mathcal{A}$ an LDA if it yields expected losses that are at least as accurate as those associated with a^* for both groups, and at least as fair, with one of these inequalities strict. In terms of the FA-dominance notion and FA-frontier reported in Definition 5, the absence of an LDA is equivalent to the fact that a^* yields expected losses with $(e_r^A(a^*), e_b^A(a^*), e_r^F(a^*) - e_r^F(a^*)) \in \mathcal{F}$. Hence, we can use the characterization in Eq. (12) to test the hypothesis that no LDA exists against the alternative that it does, as follows:

$$\mathsf{H}_0 : \min_{q\in\tilde{\mathbb{S}}^2}\left[h_{\mathcal{C}(\varepsilon^*)}(q) + h_{\mathcal{E}}(-q)\right] = 0, \tag{33}$$

$$\mathsf{H}_A : \min_{q\in\tilde{\mathbb{S}}^2}\left[h_{\mathcal{C}(\varepsilon^*)}(q) + h_{\mathcal{E}}(-q)\right] > 0, \tag{34}$$

where $\tilde{\mathbb{S}}^2 = \{q \in \mathbb{S}^2 : q_1 \geq 0, q_2 \geq 0\}$, $h_{\mathcal{C}(e^*)}(q)$ is defined in Eq. (11), and $\varepsilon^* = (e_r^A(a^*), e_b^A(a^*), e_r^F(a^*) - e_b^F(a^*))$. We refer to Eq. (33) as the LDA hypothesis.

We propose to reject the LDA hypothesis for large values of the following test statistic:

$$T_n^{\mathrm{LDA}} \equiv \sqrt{n}\left(\left[\max_{q\in\mathbb{S}^2}(q^\intercal\widehat{\varepsilon}^* - \widehat{h}_{\mathcal{E}}(q,\widehat{\mathbf{L}};\widehat{\boldsymbol{\eta}}))\right]_+ + \left[\min_{q\in\tilde{\mathbb{S}}^2}\left(h_{\mathcal{C}(\widehat{\varepsilon}^*)}(q) + \widehat{h}_{\mathcal{E}}(-q,\widehat{\mathbf{L}};\widehat{\boldsymbol{\eta}})\right)\right]_+\right), \tag{35}$$

where $\widehat{h}_{\mathcal{E}}$ and $\widehat{\varepsilon}^*$ are as in Eqs. (28) and (31), respectively. This test statistic was proposed in Liu and Molinari (2026, Section 6.2). Our contribution is deriving its asymptotic distri-

bution as well as a valid bootstrap procedure to estimate the associated critical values, in the presence of selectively observed labels.

We consider a bootstrap-based critical value $\widehat{c}_{1-\alpha}$ defined in Procedure 1 below. To simplify notation, let $\boldsymbol{\beta} = (h_{\mathcal{E}}(q), e^*) \in \ell^\infty(\mathbb{S}^2) \times \mathbb{R}^3$ denote the vector of population values and let $\widehat{\boldsymbol{\beta}} = (\widehat{h}_{\mathcal{E}}(q), \widehat{e}^*) \in \ell^\infty(\mathbb{S}^2) \times \mathbb{R}^3$ denote the vector of their estimators. We can write $T_n^{\text{LDA}} = \sqrt{n}\left(\phi(\widehat{\boldsymbol{\beta}}) - \phi(\boldsymbol{\beta})\right)$, where ϕ is defined in Eq. (49) in Appendix A.

PROCEDURE 1—Bayesian Bootstrap for the Quantiles of T_n^{LDA}:

1. *Draw* $\{W_i\}_{i=1}^n$ *i.i.d. from the exponential distribution with mean* 1 *independent of the sample* $\{(Y_i, G_i, X_i, Z_i)\}_{i=1}^n$ *and let the bootstrap analogue of* $\widehat{\boldsymbol{\beta}}$ *be* $\widetilde{\boldsymbol{\beta}} = (\widetilde{h}_{\mathcal{E}}, \widetilde{e}^*)$, *where* $\widetilde{h}_{\mathcal{E}}(q; \widetilde{\mathbf{L}}, \widehat{\boldsymbol{\eta}}) \equiv \frac{1}{n}\sum_{i=1}^n \left(\frac{W_i}{\overline{W}}\right) \zeta_i(q; \widetilde{\mathbf{L}}, \widehat{\boldsymbol{\eta}})$, $\widetilde{e}^* \equiv \frac{1}{n}\sum_{i=1}^n \left(\frac{W_i}{\overline{W}}\right) \xi_i(a; \widetilde{\mathbf{L}}, \widehat{\boldsymbol{\eta}})$, $\overline{W} = n^{-1}\sum_{i=1}^n W_i$, *and* $\widetilde{\mathbf{L}}$ *is defined as in Eqs. (1)-(2) with* μ_g *replaced by* $\widetilde{\mu}_g = n^{-1}\sum_{i=1}^n \left(\frac{W_i}{\overline{W}}\right) \mathbb{1}\{G_i = g\}$.
2. *Numerically approximate* $\phi'_{\boldsymbol{\beta}}(\cdot)$, *the directional derivative of* $\phi(\cdot)$ *at* $\boldsymbol{\beta}$, *by*

$$\widehat{\phi}'_{\boldsymbol{\beta}}(\ddot{\boldsymbol{\beta}}) = \tfrac{1}{s_n}\left(\phi\left(\widehat{\boldsymbol{\beta}} + s_n(\ddot{\boldsymbol{\beta}})\right) - \phi\left(\widehat{\boldsymbol{\beta}}\right)\right),$$

   where $\ddot{\boldsymbol{\beta}} \in \ell^\infty(\mathbb{S}^1) \times \mathbb{R}^2$ *is a candidate direction at which we evaluate* $\phi'_{\boldsymbol{\beta}}(\cdot)$ *and* s_n *is a vanishing sequence of step sizes such that* $\sqrt{n}s_n \to \infty$.
3. *Obtain* $\widehat{\phi}'_{\boldsymbol{\beta}}\big(\sqrt{n}\{\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\}\big)$ *and calculate*

$$\widehat{c}_{1-\alpha} \equiv \inf\left\{c : \mathbb{P}\left(\widehat{\phi}'_{\boldsymbol{\beta}}\big(\sqrt{n}\{\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\}\big) \le c \;\middle|\; \{(Y_i, G_i, X_i, Z_i)\}_{i=1}^n\right) \ge 1 - \alpha\right\}.$$

Let us define the LDA test by $\varphi_n^{\text{LDA}} = \mathbb{1}\{T_n^{\text{LDA}} > \widehat{c}_{1-\alpha+\kappa} + \kappa\}$ for a given significance level $\alpha \in (0, 1)$, where $\kappa > 0$ is an arbitrarily small positive constant. The constant κ can be taken equal to 10^{-6} as in Andrews and Shi (2013). We can take $\kappa = 0$ when the asymptotic distribution of T_n^{LDA} is continuous in its $(1-\alpha)$-quantile. Therefore, $\kappa > 0$ appears in the critical value used in the LDA test to handle the technicalities arising from possible discontinuity points of the asymptotic distribution of T_n^{LDA}.

The next result stablishes the consistency of the Bayesian bootstrap outlined in Procedure 1 and guarantees the asymptotic correct size of the proposed test φ_n^{LDA}.

THEOREM 4.1: *Let Assumptions 1–5 hold. Then, the Bayesian bootstrap outlined in Procedure 1 is consistent. Furthermore, if the LDA hypothesis holds, the test* φ_n^{LDA} *has asymptotically correct size,* $\limsup_{n\to\infty} \mathbb{E}[\varphi_n^{LDA}] \leq \alpha$.

Given Theorem 4.1, we can build an asymptotically valid confidence set for $\mathcal{F}$ by test inversion, similarly to Liu and Molinari (2026, Proposition 5.1).

5. CONCLUSIONS

In this paper, we consider the problem of identification and inference for the FA-frontier put forward in (LLMO), when labels (outcomes) are selectively observed. We build on the methodology developed by Liu and Molinari (2026), who assume away the selective labels problem to focus on deriving moment-inequality representations for the frontier. Selective labels challenge one's ability to (point) identify features of the conditional distribution of the true labels, including the FA-frontier. Naively computed accuracy metrics using only observed outcomes may misrepresent true performance, potentially overstating accuracy where it matters most and distorting conclusions on whether LDAs exist.

The results reported in this version of the paper rely on assuming labels are missing at random conditional on a rich set of observed covariates. We obtain point identification of the FA-frontier using inverse propensity score weighting and derive the asymptotic distribution of the test statistic that we use to characterize whether an algorithm is on the frontier. The inference method that we propose is based on a debiased machine learning estimator, which is particularly appropriate in this contexts, where high-dimensional administrative data enable flexible non-parametric modeling of the selection process and hence make the missing at random assumption more credible, but formal inference with valid confidence statements remains necessary for legally defensible determinations, making the use of methods that control the bias of the first-step nonparametric estimation crucial.

In work in progress, we dispense with the missing at random assumption and obtain a tractable characterization of the FA-frontier. We also provide a test statistic to assess existence of an LDA to a given algorithm.

REFERENCES

ADLER, ROBERT J. AND JONATHAN E. TAYLOR (2007): *Random Fields and Geometry*, Springer Monographs in Mathematics, New York: Springer. [39]

ANDREWS, DONALD W. K. (1994): "Chapter 37 Empirical process methods in econometrics," Elsevier, vol. 4 of *Handbook of Econometrics*, 2247–2294. [38]

ANDREWS, DONALD W. K. AND XIAOXIA SHI (2013): "Inference based on conditional moment inequalities," *Econometrica*, 81, 609–666. [22]

ANGWIN, JULIA, JEFF LARSON, SURYA MATTU, AND LAUREN KIRCHNER (2016): "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks." *ProPublica*, 23, 77–91. [2]

ARNOLD, DAVID, WILL DOBBIE, AND PETER HULL (2021): "Measuring racial discrimination in algorithms," in *AEA Papers and Proceedings*, vol. 111, 49–54. [2]

AUERBACH, ERIC, ANNIE LIANG, MAX TABORD-MEEHAN, AND KYOHEI OKUMURA (2024): "Testing the Fairness-Improvability of Algorithms," *arXiv preprint arXiv:2405.04816*. [9]

BAROCAS, SOLON, MORITZ HARDT, AND ARVIND NARAYANAN (2023): *Fairness and Machine Learning: Limitations and Opportunities*, MIT Press, http://www.fairmlbook.org. [5]

BERESTEANU, ARIE AND FRANCESCA MOLINARI (2008): "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76, 763–814. [4]

BERK, RICHARD, HODA HEIDARI, SHAHIN JABBARI, MATTHEW JOSEPH, MICHAEL KEARNS, JAMIE MORGENSTERN, SETH NEEL, AND AARON ROTH (2017): "A convex framework for fair regression," *arXiv preprint arXiv:1706.02409*. [5]

BERK, RICHARD, HODA HEIDARI, SHAHIN JABBARI, MICHAEL KEARNS, AND AARON ROTH (2021): "Fairness in Criminal Justice Risk Assessments: The State of the Art," *Sociological Methods & Research*, 50, 3–44. [2]

CANAY, IVAN A., GASTÓN ILLANES, AND AMILCAR VELEZ (2026): "A User's Guide for Inference in Models Defined by Moment Inequalities," *Journal of Econometrics, https://10.1016/j.jeconom.2023.105558*. [4]

CHERNOZHUKOV, VICTOR, DENIS CHETVERIKOV, MERT DEMIRER, ESTHER DUFLO, CHRISTIAN HANSEN, WHITNEY NEWEY, AND JAMES ROBINS (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21, C1–C68. [4, 17, 37, 51]

CHERNOZHUKOV, VICTOR, CHRISTIAN HANSEN, NATHAN KALLUS, MARTIN SPINDLER, AND VASILIS SYRGKANIS (2024): *Applied Causal Inference Powered by ML and AI*, Online; Version 0.1.1. [4]

CHOULDECHOVA, ALEXANDRA AND AARON ROTH (2018): "The Frontiers of Fairness in Machine Learning," *arXiv preprint arXiv:1810.08810*. [5]

CORBETT-DAVIES, SAM, JOHANN D. GAEBLER, HAMED NILFOROSHAN, RAVI SHROFF, AND SHARAD GOEL (2024): "The measure and mismeasure of fairness," *Journal of Machine Learning Research*, 24. [5]

CORBETT-DAVIES, SAM, EMMA PIERSON, AVI FELLER, SHARAD GOEL, AND AZIZ HUQ (2017): "Algorithmic Decision Making and the Cost of Fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 797–806. [3]

COSTON, AMANDA, ASHESH RAMBACHAN, AND ALEXANDRA CHOULDECHOVA (2021): "Characterizing Fairness Over the Set of Good Models Under Selective Labels," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, vol. 139 of *Proceedings of Machine Learning Research*, 2144–2155. [5]

COWGILL, BO AND CATHERINE E. TUCKER (2020): "Algorithmic Fairness and Economics," available at http://dx.doi.org/10.2139/ssrn.3361280. [2]

DWORK, CYNTHIA, MORITZ HARDT, TONIANN PITASSI, OMER REINGOLD, AND RICHARD ZEMEL (2012): "Fairness through Awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, New York, NY, USA: Association for Computing Machinery, 214–226. [5]

EUROPEAN PARLIAMENT AND COUNCIL (2024): "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence," OJ L, 2024/1689, 12.7.2024. [2, 4]

FANG, ZHENG AND ANDRES SANTOS (2019): "Inference on directionally differentiable functions," *The Review of Economic Studies*, 86, 377–412, with Online Appendix at academic.oup.com/restud/article/86/1/377/5094886#supplementary-data. [4, 38, 39]

GEBRU, TIMNIT, JAMIE MORGENSTERN, BRIANA VECCHIONE, JENNIFER WORTMAN VAUGHAN, HANNA WALLACH, HAL DAUMÉ III, AND KATE CRAWFORD (2021): "Datasheets for Datasets," *Communications of the ACM*, 64, 86–92. [2]

GILLIS, TALIA B., VITALY MEURSAULT, AND BERK USTUN (2024): "Operationalizing the Search for Less Discriminatory Alternatives in Fair Lending," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, ACM, 377–387. [3]

HECKMAN, JAMES J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161. [3]

HONG, HAN AND JESSIE LI (2018): "The numerical delta method," *Journal of Econometrics*, 206, 379–394, special issue on Advances in Econometric Theory: Essays in honor of Takeshi Amemiya. [39]

KAIDO, HIROAKI (2016): "A dual approach to inference for partially identified econometric models," *Journal of Econometrics*, 192, 269 – 290. [38]

KHAN, SHAKEEB, ELIE TAMER, AND QINGSONG YAO (2025): "Inference on High Dimensional Selective Labeling Models," *arXiv preprint arXiv:2410.18381v4*. [5, 12]

KIM, JOON SIK, JIAHAO CHEN, AND AMEET TALWALKAR (2020): "FACT: A Diagnostic for Group Fairness Trade-offs," in *Proceedings of the 37th International Conference on Machine Learning*, PMLR, vol. 119 of *Proceedings of Machine Learning Research*, 5264–5274. [3]

KLEINBERG, JON, HIMABINDU LAKKARAJU, JURE LESKOVEC, JENS LUDWIG, AND SENDHIL MULLAINATHAN (2017): "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics*, 133, 237–293. [5, 12]

LAKKARAJU, HIMABINDU, JON KLEINBERG, JURE LESKOVEC, JENS LUDWIG, AND SENDHIL MULLAINATHAN (2017): "The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables," *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [5, 12]

LIANG, ANNIE, JAY LU, XIAOSHENG MU, AND KYOHEI OKUMURA (2025): "Algorithm Design: A Fairness-Accuracy Frontier," *Journal of Political Economy, https://doi.org/10.1086/739826*. [3, 5, 6, 7, 8, 9, 10, 11, 23]

LITTLE, CAMILLE OLIVIA, MICHAEL WEYLANDT, AND GENEVERA I. ALLEN (2022): "To the Fairness Frontier and Beyond: Identifying, Quantifying, and Optimizing the Fairness-Accuracy Pareto Frontier," *arXiv preprint arXiv:2206.00074*. [5]

LIU, YIQI AND FRANCESCA MOLINARI (2026): "Inference for an Algorithmic Fairness-Accuracy Frontier," *revision in progress. Previous version available as arXiv preprint arXiv:2402.08879v2*. [3, 4, 6, 10, 11, 19, 20, 21, 23, 36]

MANSKI, CHARLES F. (1989): "Anatomy of the Selection Problem," *Journal of Human Resources*, 24, 343–360. [3, 4]

MENON, ADITYA KRISHNA AND ROBERT C. WILLIAMSON (2018): "The Cost of Fairness in Binary Classification," in *Proceedings of Machine Learning Research*, PMLR, vol. 81, 107–118. [3, 4]

MITCHELL, MARGARET, SIMONE WU, ANDREW ZALDIVAR, PARKER BARNES, LUCY VASSERMAN, BEN HUTCHINSON, ELENA SPITZER, INIOLUWA DEBORAH RAJI, AND TIMNIT GEBRU (2019): "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 220–229. [2]

MOLINARI, FRANCESCA (2020): "Microeconometrics with Partial Identification," in *Handbook of Econometrics*, Volume 7A, ed. by Steven N. Durlauf, Lars Peter Hansen, James J. Heckman, and Rosa L. Matzkin, Amsterdam: Elsevier, 355–486. [4]

OBERMEYER, ZIAD, BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN (2019): "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, 366, 447–453. [2]

RAJI, INIOLUWA DEBORAH, ANDREW SMART, REBECCA N. WHITE, MARGARET MITCHELL, TIMNIT GEBRU, BEN HUTCHINSON, JAMILA SMITH-LOUD, DANIEL THERON, AND PARKER BARNES (2020): "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*)*, 33–44. [2]

RAMBACHAN, ASHESH, AMANDA COSTON, AND EDWARD H. KENNEDY (2025): "Robust Design and Evaluation of Predictive Algorithms under Confounding," *arXiv preprint arXiv:2212.09844v6*. [5, 12]

RAO, M. M. AND R. J. SWIFT (2006): *Probability Theory With Applications*, vol. 582 of *Mathematics and Its Applications*, New York: Springer-Verlag. [41]

ROMANO, JOSEPH P AND AZEEM M SHAIKH (2012): "On the uniform asymptotic validity of subsampling and the bootstrap," . [38, 40]

SCHNEIDER, ROLF (1993): *Convex Bodies: The Brunn-Minkowski Theory*, Encyclopedia of Mathematics and its Applications, Cambridge University Press. [29, 43]

SEMENOVA, VIRA (2023): "Debiased machine learning of set-identified linear models," *Journal of Econometrics*, 235, 1725–1746. [4]

U.S. CONGRESS (1991): "Civil Rights Act of 1991, Pub. L. No. 102-166, 105 Stat. 1071," Codified at 42 U.S.C. § 2000e et seq. Section 2000e-2(k) establishes disparate impact framework and alternative employment practice requirement. [3, 4]

VAN DER VAART, AAD W. AND JON A. WELLNER (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Series in Statistics, Springer New York. [35, 37, 38, 39]

WEI, DENNIS AND MARC NIETHAMMER (2022): "Estimating and Improving Fairness with Adversarial Learning," in *arXiv preprint arXiv:2103.04243*. [5]

APPENDIX A: PROOFS OF THE MAIN RESULTS

A.1. *Proofs for Section 2*

PROOF OF THEOREM 2.1: Recall $\mathcal{E} = \left\{ \left(e_r^A(a),\, e_b^A(a),\, e_r^F(a) - e_b^F(a) \right) : a \in \mathcal{A} \right\}$, so

$$h_{\mathcal{E}}(q) = \sup_{a \in \mathcal{A}} q^\intercal \left(e_r^A(a),\, e_b^A(a),\, e_r^F(a) - e_b^F(a) \right).$$

Using the representation $\left(e_r^A(a),\, e_b^A(a),\, e_r^F(a) - e_b^F(a) \right) = \mathbb{E}^*[\boldsymbol{\theta}_0(X)] + \mathbb{E}^*\left[a(X)\Delta\boldsymbol{\theta}(X)\right]$, we obtain $q^\intercal \left(e_r^A(a),\, e_b^A(a),\, e_r^F(a) - e_b^F(a) \right) = \mathbb{E}^*\left[q^\intercal\boldsymbol{\theta}_0(X)\right] + \mathbb{E}^*\left[a(X) q^\intercal\Delta\boldsymbol{\theta}(X)\right]$. Hence,

$$h_{\mathcal{E}}(q) = \mathbb{E}^*\left[q^\intercal\boldsymbol{\theta}_0(X)\right] + \sup_{a \in \mathcal{A}} \mathbb{E}^*\left[a(X) q^\intercal\Delta\boldsymbol{\theta}(X)\right].$$

Because $a(X) \in [0,1]$ and $q^\intercal\Delta\boldsymbol{\theta}(X)$ is $\sigma(X)$-measurable, the pointwise maximizer is

$$a^q(X) \in \arg\max_{t \in [0,1]} t\, q^\intercal\Delta\boldsymbol{\theta}(X) = \begin{cases} 1, & q^\intercal\Delta\boldsymbol{\theta}(X) > 0, \\ 0, & q^\intercal\Delta\boldsymbol{\theta}(X) < 0, \end{cases}$$

with arbitrary tie-breaking on $\{q^\intercal\Delta\boldsymbol{\theta}(X) = 0\}$. Therefore,

$$\sup_{a \in \mathcal{A}} \mathbb{E}^*\left[a(X) q^\intercal\Delta\boldsymbol{\theta}(X)\right] = \mathbb{E}^*\left[\left(q^\intercal\Delta\boldsymbol{\theta}(X)\right)_+\right].$$

Next, note that $\mathbb{1}\{q^\intercal\Delta\boldsymbol{\theta}(X) > 0\}$ is $\sigma(X)$-measurable. Therefore, by iterated expectations,

$$\mathbb{E}^*\left[q^\intercal \mathbf{L}_0 + q^\intercal(\mathbf{L}_1 - \mathbf{L}_0)\mathbb{1}\{q^\intercal \Delta\boldsymbol{\theta}(X) > 0\}\right]$$
$$= \mathbb{E}^*\left[q^\intercal \boldsymbol{\theta}_0(X) + q^\intercal \Delta\boldsymbol{\theta}(X)\mathbb{1}\{q^\intercal \Delta\boldsymbol{\theta}(X) > 0\}\right].$$

This proves (8).

To prove (9), fix $q \in \mathbb{S}^2$ and let $v \in \mathbb{R}^3$ be arbitrary; denote $u_+ \equiv \max\{u, 0\}$. For $t \in \mathbb{R}$ such that $q + tv \neq 0$, using the support-function formula, we have

$$\frac{h_{\mathcal{E}}(q+tv) - h_{\mathcal{E}}(q)}{t} = \mathbb{E}^*\left[v^\intercal \boldsymbol{\theta}_0(X) + \frac{((q+tv)^\intercal \Delta\boldsymbol{\theta}(X))_+ - (q^\intercal \Delta\boldsymbol{\theta}(X))_+}{t}\right].$$

Hence it is enough to study the difference quotient of the map $u \mapsto u_+$.

For each realization $X = x$ such that $q^\intercal \Delta\boldsymbol{\theta}(x) \neq 0$, the function $t \mapsto ((q + tv)^\intercal \Delta\boldsymbol{\theta}(x))_+$ is differentiable at $t = 0$, with derivative $v^\intercal \Delta\boldsymbol{\theta}(x)\mathbb{1}\{q^\intercal \Delta\boldsymbol{\theta}(x) > 0\}$. Therefore, pointwise on the event $\{q^\intercal \Delta\boldsymbol{\theta}(X) \neq 0\}$, as $t \to 0$,

$$\frac{((q+tv)^\intercal \Delta\boldsymbol{\theta}(X))_+ - (q^\intercal \Delta\boldsymbol{\theta}(X))_+}{t} \to v^\intercal \Delta\boldsymbol{\theta}(X)\mathbb{1}\{q^\intercal \Delta\boldsymbol{\theta}(X) > 0\}. \tag{36}$$

Assumption 2 implies $\mathbb{P}^*\big(q^\intercal \Delta\boldsymbol{\theta}(X) = 0\big) = 0$ for every $q \in \mathbb{S}^2$, because for every $\delta > 0$,

$$\mathbb{P}^*\left(q^\intercal \Delta\boldsymbol{\theta}(X) = 0\right) \leq \sup_{\tilde{q} \in \mathbb{S}^2} \mathbb{P}^*\big(|\tilde{q}^\intercal \Delta\boldsymbol{\theta}(X)| \leq \delta\big) \leq C\delta^m$$

and letting $\delta \downarrow 0$ yields the claim. Thus, the convergence in (36) holds $\mathbb{P}^*$-almost surely.

Next we obtain an integrable dominating function. Since $u \mapsto u_+$ is Lipschitz with constant one,

$$\left|\frac{((q+tv)^\intercal \Delta\boldsymbol{\theta}(X))_+ - (q^\intercal \Delta\boldsymbol{\theta}(X))_+}{t}\right| \leq |v^\intercal \Delta\boldsymbol{\theta}(X)|$$

for all $t \neq 0$. Therefore,

$$\left|v^\intercal \boldsymbol{\theta}_0(X) + \frac{((q+tv)^\intercal \Delta\boldsymbol{\theta}(X))_+ - (q^\intercal \Delta\boldsymbol{\theta}(X))_+}{t}\right| \leq |v^\intercal \boldsymbol{\theta}_0(X)| + |v^\intercal \Delta\boldsymbol{\theta}(X)|$$

By Assumption 1, the random vectors $\boldsymbol{\theta}_0(X)$ and $\Delta\boldsymbol{\theta}(X)$ are integrable, so the right-hand side is integrable. Hence, by dominated convergence,

$$\lim_{t \to 0} \frac{h_{\mathcal{E}}(q+tv) - h_{\mathcal{E}}(q)}{t} = v^\intercal \mathbb{E}^*\left[\boldsymbol{\theta}_0(X) + \Delta\boldsymbol{\theta}(X)\mathbb{1}\{q^\intercal \Delta\boldsymbol{\theta}(X) > 0\}\right].$$

Thus the directional derivative of $h_{\mathcal{E}}$ at q in direction v exists and is linear in v. It follows that $h_{\mathcal{E}}$ is differentiable at q, with gradient

$$\nabla_q h_{\mathcal{E}}(q) = \mathbb{E}^* \left[\boldsymbol{\theta}_0(X) + \Delta\boldsymbol{\theta}(X) \mathbb{1}\{q^\intercal \Delta\boldsymbol{\theta}(X) > 0\} \right].$$

Using again the law of iterated expectations yields (9).

Finally, by Schneider (1993, Corollary 1.7.3), $\mathcal{E}$ is strictly convex because by (9) its support set is a singleton in each direction $q \in \mathbb{S}^2$. *Q.E.D.*

PROOF OF THEOREM 2.2: Recall $\widetilde{\mathbb{S}}^2 \equiv \{q \in \mathbb{S}^2 : q_1 \geq 0,\ q_2 \geq 0\}$ and that under Assumptions 1-2, $\mathcal{E}$ has a nonempty interior, is compact, and is strictly convex.

Step 1: (i) $\iff$ (ii). By definition,

$$\mathcal{C}(\varepsilon^*) = \left\{ \varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3) \in \mathbb{R}^3 : \varepsilon_1 \leq \varepsilon_1^*,\ \varepsilon_2 \leq \varepsilon_2^*,\ |\varepsilon_3| \leq |\varepsilon_3^*| \right\}.$$

Hence $\mathcal{E} \cap \mathcal{C}(\varepsilon^*) = \left\{ \varepsilon \in \mathcal{E} : \varepsilon_1 \leq \varepsilon_1^*,\ \varepsilon_2 \leq \varepsilon_2^*,\ |\varepsilon_3| \leq |\varepsilon_3^*| \right\}$. Next, recall that $\varepsilon^* \in \mathcal{F}$ if and only if there does not exist $\varepsilon \in \mathcal{E}$ with $\varepsilon_1 \leq \varepsilon_1^*$, $\varepsilon_2 \leq \varepsilon_2^*$, and $|\varepsilon_3| \leq |\varepsilon_3^*|$, with at least one strict inequality. Since $\varepsilon^* \in \mathcal{E} \cap \mathcal{C}(\varepsilon^*)$, this is equivalent to $\mathcal{E} \cap \mathcal{C}(\varepsilon^*) = \{\varepsilon^*\}$.

Step 2: (ii) $\iff$ (iv). We begin showing $(ii) \Rightarrow (iv)$. Assume $\mathcal{E} \cap \mathcal{C}(\varepsilon^*) = \{\varepsilon^*\}$. We prove first that $|\varepsilon_3^*| = d((\varepsilon_1^*, \varepsilon_2^*))$. If not, then $d((\varepsilon_1^*, \varepsilon_2^*)) < |\varepsilon_3^*|$, and by definition of d, there exists $\tilde{\varepsilon}_3 \in \mathbb{R}$ such that $(\varepsilon_1^*, \varepsilon_2^*, \tilde{\varepsilon}_3) \in \mathcal{E}$ and $|\tilde{\varepsilon}_3| < |\varepsilon_3^*|$. But then $(\varepsilon_1^*, \varepsilon_2^*, \tilde{\varepsilon}_3) \in \mathcal{E} \cap \mathcal{C}(\varepsilon^*)$ and this point is distinct from ε^*, contradicting $\mathcal{E} \cap \mathcal{C}(\varepsilon^*) = \{\varepsilon^*\}$. Hence, $|\varepsilon_3^*| = d((\varepsilon_1^*, \varepsilon_2^*))$. We next show that $\pi(\varepsilon^*) = (\varepsilon_1^*, \varepsilon_2^*) \in \mathcal{F}^A$. Suppose not. Then there exists $(c_r, c_b) \in \mathcal{E}^A$ such that $c_r \leq \varepsilon_1^*$, $c_b \leq \varepsilon_2^*$, and $d((c_r, c_b)) \leq d((\varepsilon_1^*, \varepsilon_2^*))$, with at least one strict inequality. As $(c_r, c_b) \in \mathcal{E}^A$, by definition of d there exists $\tilde{\varepsilon}_3 \in \mathbb{R}$ such that $(c_r, c_b, \tilde{\varepsilon}_3) \in \mathcal{E}$ and $|\tilde{\varepsilon}_3| = d((c_r, c_b))$. Using the equality already proved, $d((\varepsilon_1^*, \varepsilon_2^*)) = |\varepsilon_3^*|$, we obtain $c_r \leq \varepsilon_1^*$, $c_b \leq \varepsilon_2^*$, and $|\tilde{\varepsilon}_3| \leq |\varepsilon_3^*|$, with at least one strict inequality. Thus $(c_r, c_b, \tilde{\varepsilon}_3) \in \mathcal{E} \cap \mathcal{C}(\varepsilon^*)$ with $(c_r, c_b, \tilde{\varepsilon}_3) \neq \varepsilon^*$, contradicting $\mathcal{E} \cap \mathcal{C}(\varepsilon^*) = \{\varepsilon^*\}$. Hence, $\pi(\varepsilon^*) \in \mathcal{F}^A$, proving $(ii) \Rightarrow (iv)$.

We now prove $(iv) \Rightarrow (ii)$. Assume $\pi(\varepsilon^*) = (\varepsilon_1^*, \varepsilon_2^*) \in \mathcal{F}^A$ and $|\varepsilon_3^*| = d((\varepsilon_1^*, \varepsilon_2^*))$. Let $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3) \in \mathcal{E} \cap \mathcal{C}(\varepsilon^*)$. Then $\varepsilon_1 \leq \varepsilon_1^*$, $\varepsilon_2 \leq \varepsilon_2^*$, and $|\varepsilon_3| \leq |\varepsilon_3^*|$. Since $\varepsilon \in \mathcal{E}$, we have

$$d((\varepsilon_1, \varepsilon_2)) \leq |\varepsilon_3| \leq |\varepsilon_3^*| = d((\varepsilon_1^*, \varepsilon_2^*)). \tag{37}$$

Thus, $(\varepsilon_1, \varepsilon_2) \in \mathcal{E}^A$ weakly improves on $(\varepsilon_1^*, \varepsilon_2^*)$ in both group-accuracy losses and in the fairness index d. As $(\varepsilon_1^*, \varepsilon_2^*) \in \mathcal{F}^A$, it follows that $(\varepsilon_1, \varepsilon_2) = (\varepsilon_1^*, \varepsilon_2^*)$ (else we would contradict $(\varepsilon_1^*, \varepsilon_2^*) \in \mathcal{F}^A$). Substituting back $(\varepsilon_1^*, \varepsilon_2^*)$ for $(\varepsilon_1, \varepsilon_2)$ in (37) yields $|\varepsilon_3| = |\varepsilon_3^*|$. If $\varepsilon_3 \neq \varepsilon_3^*$, then necessarily $\varepsilon_3 = -\varepsilon_3^*$. If $\varepsilon_3^* = 0$, this is impossible, so $\varepsilon_3 = \varepsilon_3^*$. If $\varepsilon_3^* \neq 0$, by convexity of $\mathcal{E}$ the midpoint $\bar{\varepsilon} \equiv \frac{\varepsilon + \varepsilon^*}{2} = (\varepsilon_1^*, \varepsilon_2^*, 0)$ belongs to $\mathcal{E}$. Therefore, $d((\varepsilon_1^*, \varepsilon_2^*)) \leq 0$, which implies $d((\varepsilon_1^*, \varepsilon_2^*)) = 0$. But then $|\varepsilon_3^*| = d((\varepsilon_1^*, \varepsilon_2^*)) = 0$, contradicting $\varepsilon_3^* \neq 0$. Hence $\varepsilon_3 = \varepsilon_3^*$, so $\varepsilon = \varepsilon^*$. We conclude that $\mathcal{E} \cap \mathcal{C}(\varepsilon^*) = \{\varepsilon^*\}$, proving $(iv) \Rightarrow (ii)$.

Step 3: (ii) $\Rightarrow$ (iii). Assume $\mathcal{E} \cap \mathcal{C}(\varepsilon^*) = \{\varepsilon^*\}$. As $\mathcal{E}$ is compact and convex with nonempty interior, and $\mathcal{C}(\varepsilon^*)$ is closed and convex, the two sets admit a common supporting hyperplane at their unique common point. Hence, there exists $q \neq 0$ and $\alpha \in \mathbb{R}$ such that

$$q^\intercal \varepsilon \geq \alpha \geq q^\intercal c \qquad \forall \varepsilon \in \mathcal{E},\ \forall c \in \mathcal{C}(\varepsilon^*).$$

As $\varepsilon^* \in \mathcal{E} \cap \mathcal{C}(\varepsilon^*)$, necessarily $\alpha = q^\intercal \varepsilon^*$, hence $\inf_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon = \sup_{c \in \mathcal{C}(\varepsilon^*)} q^\intercal c$. Equivalently, $-h_{\mathcal{E}}(-q) = h_{\mathcal{C}(\varepsilon^*)}(q)$. Since $h_{\mathcal{C}(\varepsilon^*)}(q) < +\infty$, we must have $q_1 \geq 0$ and $q_2 \geq 0$, because $\mathcal{C}(\varepsilon^*)$ is unbounded in the negative first- and second-coordinate directions. Since $q \neq 0$, positive homogeneity of support functions implies that, after replacing q by $q/\|q\|$, we may assume $q \in \widetilde{\mathbb{S}}^2$ while preserving the equality $h_{\mathcal{C}(\varepsilon^*)}(q) + h_{\mathcal{E}}(-q) = 0$. Since $\varepsilon^* \in \mathcal{E} \cap \mathcal{C}(\varepsilon^*)$, for every $q \in \widetilde{\mathbb{S}}^2$, $h_{\mathcal{C}(\varepsilon^*)}(q) \geq q^\intercal \varepsilon^*$ and $h_{\mathcal{E}}(-q) \geq -q^\intercal \varepsilon^*$, and hence $h_{\mathcal{C}(\varepsilon^*)}(q) + h_{\mathcal{E}}(-q) \geq 0$. Therefore, $\min_{q \in \widetilde{\mathbb{S}}^2} \left(h_{\mathcal{C}(\varepsilon^*)}(q) + h_{\mathcal{E}}(-q) \right) = 0$, proving $(ii) \Rightarrow (iii)$.

Step 4: (iii) $\Rightarrow$ (ii). Assume $\min_{q \in \widetilde{\mathbb{S}}^2} \left(h_{\mathcal{C}(\varepsilon^*)}(q) + h_{\mathcal{E}}(-q) \right) = 0$. As $\widetilde{\mathbb{S}}^2$ is compact, $\mathcal{E}$ is compact, and $h_{\mathcal{C}(\varepsilon^*)}$ is finite and continuous on $\widetilde{\mathbb{S}}^2$, the minimum is attained at some $\bar{q} \in \widetilde{\mathbb{S}}^2$. Thus, $\sup_{c \in \mathcal{C}(\varepsilon^*)} \bar{q}^\intercal c = \inf_{\varepsilon \in \mathcal{E}} \bar{q}^\intercal \varepsilon$. Let $\tilde{\varepsilon} \in \mathcal{E} \cap \mathcal{C}(\varepsilon^*)$. Then

$$\bar{q}^\intercal \tilde{\varepsilon} \leq \sup_{c \in \mathcal{C}(\varepsilon^*)} \bar{q}^\intercal c = \inf_{\varepsilon \in \mathcal{E}} \bar{q}^\intercal \varepsilon \leq \bar{q}^\intercal \tilde{\varepsilon}.$$

Hence, every $\tilde{\varepsilon} \in \mathcal{E} \cap \mathcal{C}(\varepsilon^*)$ satisfies $\bar{q}^\intercal \tilde{\varepsilon} = \inf_{\varepsilon \in \mathcal{E}} \bar{q}^\intercal \varepsilon$. Therefore,

$$\mathcal{E} \cap \mathcal{C}(\varepsilon^*) \subset \left\{ \varepsilon \in \mathcal{E} : \bar{q}^\intercal \varepsilon = \inf_{\tilde{\varepsilon} \in \mathcal{E}} \bar{q}^\intercal \tilde{\varepsilon} \right\},$$

that is, the intersection is contained in the support set of $\mathcal{E}$ in direction $-\bar{q}$. Since $\mathcal{E}$ is strictly convex, every support set of $\mathcal{E}$ is a singleton. Because $\varepsilon^* \in \mathcal{E} \cap \mathcal{C}(\varepsilon^*)$, the above inclusion implies $\mathcal{E} \cap \mathcal{C}(\varepsilon^*) = \{\varepsilon^*\}$, proving $(iii) \Rightarrow (ii)$.

Combining Steps 1-4, we obtain $(i) \iff (ii) \iff (iii) \iff (iv)$. *Q.E.D.*

PROOF OF THEOREM 2.3: Let $RB(\alpha) \equiv \arg\min_{\varepsilon \in \mathcal{E}} \left[\alpha\varepsilon_1 + (1-\alpha)\varepsilon_2\right]$, $\alpha \in [0,1]$ and $RB([0,1]) \equiv \{RB(\alpha) : \alpha \in [0,1]\}$. By Lemma B.4, $\mathcal{PF} = RB([0,1])$. By strict convexity of $\mathcal{E}$, for any $q(\alpha)$, $RB(\alpha) = \arg\min_{\eta \in \mathcal{E}} q(\alpha)^\intercal \eta$ is unique, so $\varepsilon^* \in RB([0,1])$ if and only if there exists $\alpha \in [0,1]$ such that $\varepsilon^* = RB(\alpha)$.

As $RB([0,1]) = \mathcal{PF}$ by Lemma B.4, it follows that $RB([0,1]) \subseteq \mathcal{F}$. We then prove that the reverse inclusion holds if and only if the condition in Theorem 2.3 holds.

Assume first that $\forall \varepsilon^* \in \mathcal{F}$, there exists $\alpha \in [0,1]$ such that $\varepsilon^* = \arg\min_{\eta \in \mathcal{E}} q(\alpha)^\intercal \eta$. Then $\varepsilon^* \in RB([0,1])$. Hence $\mathcal{F} \subseteq RB([0,1])$ and therefore $\mathcal{F} = RB([0,1]) = \mathcal{PF}$.

Conversely, assume that $RB([0,1]) = \mathcal{F}$. Let $\varepsilon^* \in \mathcal{F}$. Then $\varepsilon^* \in RB([0,1])$, so $\varepsilon^* = RB(\alpha)$ for some $\alpha \in [0,1]$. Then the argument in the first paragraph of this proof yields $\varepsilon^* = \arg\min_{\eta \in \mathcal{E}} q(\alpha)^\intercal \eta$. Since $\varepsilon^* \in \mathcal{F}$ was arbitrary, we obtain that for all $\varepsilon^* \in \mathcal{F}$ there exists $\alpha \in [0,1]$ such that $\varepsilon^* = \arg\min_{\eta \in \mathcal{E}} q(\alpha)^\intercal \eta$, proving the theorem. *Q.E.D.*

A.2. *Proofs for Section 3*

A.2.1. *Proofs for Section 3.1*

For $\lambda \in \Lambda$, let $\boldsymbol{\theta}_d(x;\lambda)$ denote the completion-dependent analogue of $\boldsymbol{\theta}_d(x)$. Under the special losses considered here,

$$
\begin{aligned}
L_0^{g,A} &= \tfrac{Y_i^* \mathbb{1}\{G=g\}}{\mu_g} \\
L_1^{g,A} &= \tfrac{(1-Y_i^*)\mathbb{1}\{G=g\}}{\mu_g} \\
L_0^{r,F} - L_0^{b,F} &= 0 \\
L_1^{r,F} - L_1^{b,F} &= \tfrac{\mathbb{1}\{G=r\}}{\mu_r} - \tfrac{\mathbb{1}\{G=b\}}{\mu_b}
\end{aligned}
$$

The next lemma gives the λ-dependent representations used in Section 3.1.

LEMMA A.1: *For every $\lambda \in \Lambda$ and every $x \in \mathcal{X}$,*

$$\boldsymbol{\theta}_0(x;\lambda) = A_0(x) + B(x)\lambda(x), \qquad \boldsymbol{\theta}_1(x;\lambda) = A_1(x) - B(x)\lambda(x). \tag{38}$$

Consequently,

$$h_{\mathcal{E}}(q;\lambda) = \mathbb{E}[\max\left\{q^{\intercal}(A_0(X) + B(X)\lambda(X)), q^{\intercal}(A_1(X) - B(X)\lambda(X))\right\}], \tag{39}$$

$$h_{\mathcal{C}(\varepsilon^*(a;\lambda))}(q) = C(q;a) + \mathbb{E}[(1-2a(X))q^{\intercal}B(X)\lambda(X)], \qquad q \in \widetilde{\mathbb{S}}^2, \tag{40}$$

where

$$C(q;a) \equiv \sum_{j=1}^{2} q_j \mathbb{E}[a(X)A_{1,j}(X) + (1-a(X))A_{0,j}(X)] + |q_3|\,|\mathbb{E}[a(X)A_{1,3}(X)]|. \tag{41}$$

PROOF: Fix $g \in \{r, b\}$. Using $Y = ZY^*$ and the definition of $\lambda_g(x)$,

$$\begin{aligned}
\theta_0^{g,A}(x;\lambda) &= \mathbb{E}^*\left[\frac{Y^*\mathbb{1}\{G=g\}}{\mu_g}\middle| X = x\right] \\
&= \mathbb{E}\left[\frac{ZY\mathbb{1}\{G=g\}}{\mu_g}\middle| X = x\right] + \mathbb{E}\left[\frac{(1-Z)Y^*\mathbb{1}\{G=g\}}{\mu_g}\middle| X = x\right] \\
&= \mathbb{E}\left[\frac{ZY\mathbb{1}\{G=g\}}{\mu_g}\middle| X = x\right] + \lambda_g(x)\,\mathbb{E}\left[\frac{(1-Z)\mathbb{1}\{G=g\}}{\mu_g}\middle| X = x\right],
\end{aligned}$$

which is the gth accuracy coordinate of $A_0(x) + B(x)\lambda(x)$. Likewise,

$$\begin{aligned}
\theta_1^{g,A}(x;\lambda) &= \mathbb{E}^*\left[\frac{(1-Y^*)\mathbb{1}\{G=g\}}{\mu_g}\middle| X = x\right] \\
&= \mathbb{E}\left[\frac{(Z(1-Y)+(1-Z))\mathbb{1}\{G=g\}}{\mu_g}\middle| X = x\right] - \lambda_g(x)\,\mathbb{E}\left[\frac{(1-Z)\mathbb{1}\{G=g\}}{\mu_g}\middle| X = x\right],
\end{aligned}$$

which is the gth accuracy coordinate of $A_1(x) - B(x)\lambda(x)$. For the fairness coordinate,

$$\theta_0^F(x;\lambda) = 0, \qquad \theta_1^F(x;\lambda) = \mathbb{E}\left[\frac{\mathbb{1}\{G=r\}}{\mu_r} - \frac{\mathbb{1}\{G=b\}}{\mu_b}\middle| X = x\right],$$

which does not depend on λ. This proves (38). Equation (39) now follows directly from Theorem 2.1 applied to the λ-dependent pair $(\boldsymbol{\theta}_0(\cdot;\lambda), \boldsymbol{\theta}_1(\cdot;\lambda))$.

Next fix an algorithm a. By construction,

$$\varepsilon^*(a;\lambda) = \mathbb{E}[a(X)\boldsymbol{\theta}_1(X;\lambda) + (1-a(X))\boldsymbol{\theta}_0(X;\lambda)]\,.$$

For $j \in \{1,2\}$, using (38),

$$\varepsilon^*_j(a;\lambda) = \mathbb{E}[a(X)A_{1,j}(X) + (1-a(X))A_{0,j}(X)] + \mathbb{E}[(1-2a(X))B_{jj}(X)\lambda_j(X)]\,.$$

For the third coordinate,

$$\varepsilon^*_3(a) = \mathbb{E}[a(X)A_{1,3}(X)]\,,$$

so it is point identified. Equation (11) then gives, for $q \in \widetilde{\mathbb{S}}^2$,

$$\begin{aligned} h_{\mathcal{C}(\varepsilon^*(a;\lambda))}(q) &= q_1\varepsilon^*_1(a;\lambda) + q_2\varepsilon^*_2(a;\lambda) + |q_3|\,|\varepsilon^*_3(a)| \\ &= C(q;a) + \mathbb{E}[(1-2a(X))q^\intercal B(X)\lambda(X)]\,, \end{aligned}$$

which is (40). *Q.E.D.*

PROOF OF THEOREM 3.1: By Theorem 2.2, algorithm a^* yields $\varepsilon^* \in \mathcal{F}$ if and only if

$$\min_{q\in\tilde{\mathbb{S}}^2} \left[h_{\mathcal{C}(\varepsilon^*)}(q) + h_{\mathcal{E}}(-q)\right] = 0.$$

Given the definition of $J_d(\lambda;q,a)$, $d=0,1$, in Theorem 3.1, this is equivalent to

$$\min_{q\in\tilde{\mathbb{S}}^2} \mathbb{E}^*\left[\max\left\{J_1(\lambda^*(X);q,a^*), J_0(\lambda^*(X);q,a^*)\right\}\right] = 0 \tag{42}$$

Since $\lambda^*(X)$ is not identified by the data, we say that a^* is on the fairness-accuracy frontier for some distribution consistent with the data if we can find a $\lambda^{(a^*)}(\cdot) \in \Lambda$ such that (42) holds with $\lambda^*(X) = \lambda^{(a^*)}(X)$. We can write the existence equivalently as follows:

$$\min_{\lambda(\cdot)\in\Lambda} \min_{q\in\tilde{\mathbb{S}}^2} \mathbb{E}\left[\max\left\{J_1(\lambda(X);q,a^*), J_0(\lambda(X);q,a^*)\right\}\right] = 0 \tag{43}$$

To simplify notation we use $J_d(\lambda)$ instead of $J_d(\lambda(X);q,a^*)$. Since $J_d(\lambda)$ is linear in λ, we write $J_d(\lambda) = \tilde{A}_d + \tilde{B}_d^\intercal\lambda$, where $\tilde{B}_d = (\tilde{B}_d^r, \tilde{B}_d^b)^\intercal$. We use the previous notation to write

the left-hand side of (43) as follows

$$
\begin{aligned}
&\overset{(1)}{=} \min_{q\in\tilde{\mathbb{S}}^2} \min_{\lambda(\cdot)\in\Lambda} \mathbb{E}^* \left[\max_{t\in[0,1]} \{t\ J_1(\lambda) + (1-t)\ J_0(\lambda)\} \right] \\
&\overset{(2)}{=} \min_{q\in\tilde{\mathbb{S}}^2} \mathbb{E} \left[\min_{\lambda(\cdot)\in\Lambda} \max_{t\in[0,1]} \{t\ J_1(\lambda) + (1-t)\ J_0(\lambda)\} \right] \\
&\overset{(3)}{=} \min_{q\in\tilde{\mathbb{S}}^2} \mathbb{E} \left[\max_{t\in[0,1]} \min_{\lambda(\cdot)\in\Lambda} \left\{ t\left(\tilde{A}_1 + \tilde{B}_1^\intercal \lambda\right) + (1-t)\left(\tilde{A}_0 + \tilde{B}_0^\intercal \lambda\right) \right\} \right] \\
&\overset{(4)}{=} \min_{q\in\tilde{\mathbb{S}}^2} \mathbb{E} \left[\max_{t\in[0,1]} \left\{ t\tilde{A}_1 + (1-t)\tilde{A}_0 + \min\{0, t\tilde{B}_1^r + (1-t)\tilde{B}_0^r\} + \min\{0, t\tilde{B}_1^b + (1-t)\tilde{B}_0^b\} \right\} \right] \\
&\overset{(5)}{=} \min_{q\in\tilde{\mathbb{S}}^2} \mathbb{E} \left[\max \left\{ \tilde{A}_0 + \min\{0, \tilde{B}_0^r\} + \min\{0, \tilde{B}_0^b\},\ a^*(X)\tilde{A}_1 + (1-a^*(X))\tilde{A}_0,\ \tilde{A}_1 \right\} \right]
\end{aligned}
\tag{44}
$$

where (1) holds because $\max\{a,b\} = \max_{t\in[0,1]}\{ta + (1-t)b\}$, (2) holds by the measurability selection theorem, (3) holds by the minimax theorem since the objective function is linear in t and λ and both domains are convex, (4) by solving in $\lambda(X) \in [0,1]^2$ conditional on X, and (5) holds by claims 1 and 2 below. Finally, Eq. (44) is sufficient to conclude the proof of the proposition since $\tilde{A}_0 + \min\{0, \tilde{B}_0^r\} + \min\{0, \tilde{B}_0^b\} = J_0(\lambda_1; q, a^*)$, $\tilde{A}_0 = J_0(\lambda_0; q, a^*)$, and $\tilde{A}_1 = J_1(\lambda_0; q, a^*)$.

Claim 1: $t\tilde{B}_1^g + (1-t)\tilde{B}_0^g \leq 0$ if $t \leq t^* = a^*(X)$. Recall that $\tilde{B}_d = (\tilde{B}_d^r, \tilde{B}_d^b)^\intercal$ and

$$
\begin{aligned}
\tilde{B}_0 &= -2a^*(X) q^\intercal B(X)\lambda(X) \\
\tilde{B}_1 &= 2(1-a^*(X)) q^\intercal B(X)\lambda(X)\ .
\end{aligned}
$$

Therefore, $t\tilde{B}_1 + (1-t)\tilde{B}_0 = 2\left(t - a^*(X)\right) q^\intercal B(X)\lambda(X)$, which implies the claim since $q_1 \geq 0$, $q_2 \geq 0$, and

$$
B(X) = \begin{bmatrix} \frac{(1-Z)\mathbb{1}\{G=r\}}{\mu_r} & 0 \\ 0 & \frac{(1-Z)\mathbb{1}\{G=b\}}{\mu_b} \\ 0 & 0 \end{bmatrix}\ .
$$

Claim 2: $t\tilde{B}_1^g + (1-t)\tilde{B}_0^g \geq 0$ if $t \geq t^* = a^*(X)$. Similar to claim 1.

Claim 3: $J(t)$ is maximized at $t=0$, $a^*(X)$, and 1, where

$$J(t) \equiv t\tilde{A}_1 + (1-t)\tilde{A}_0 + \min\{0, t\tilde{B}_1^r + (1-t)\tilde{B}_0^r\} + \min\{0, t\tilde{B}_1^b + (1-t)\tilde{B}_0^b\} \ .$$

Furthermore, $J(a^*(X)) = a^*(X)\tilde{A}_1 + (1-a^*(X))\tilde{A}_0$ and $J(1) = \tilde{A}_1$. The proof of this claim follows by linearity on t and claims 1 and 2. *Q.E.D.*

A.2.2. *Proofs for Section 3.2*

PROOF OF LEMMA 3.1: Since $\boldsymbol{\pi}(X) \in (0,1)$ a.s., the ratio $Z\mathbf{L}_d/\boldsymbol{\pi}(X)$ is well-defined. Applying the law of iterated expectations,

$$\mathbb{E}\left[\frac{Z\mathbf{L}_d}{\boldsymbol{\pi}(X)}\middle| X\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{Z\mathbf{L}_d}{\boldsymbol{\pi}(X)}\middle| X, Y^*, G\right]\middle| X\right] = \mathbb{E}\left[\frac{\mathbf{L}_d}{\boldsymbol{\pi}(X)}\mathbb{E}[Z|X,Y^*,G]\middle| X\right].$$

By Assumption 3, $\mathbb{E}[Z|X,Y^*,G] = \boldsymbol{\pi}(X)$ a.s., and hence $\frac{\mathbf{L}_d}{\boldsymbol{\pi}(X)}\mathbb{E}[Z|X,Y^*,G] = \mathbf{L}_d$, a.s., and the characterization of $\boldsymbol{\theta}_d(X)$ in Eq. (22) follows. Repeating the same argument and using Eq. (8) yields the characterizations of $h_{\mathcal{E}}(q)$ and $e_g^{\iota}(a)$ in Eqs. (23)-(24). *Q.E.D.*

PROOF OF THEOREM 3.2: The proof has two parts. We first show in part 1 that

$$\sup_{q\in\mathbb{S}^2} |\sqrt{n}(\widehat{h}_{\mathcal{E}}(q, \widehat{\mathbf{L}}; \widehat{\boldsymbol{\eta}}) - h_{\mathcal{E}}(q)) - \mathbb{G}_n[\zeta_i^*(q; \mathbf{L}, \boldsymbol{\eta})]| = o_p(1) \ . \tag{45}$$

We then use van der Vaart and Wellner (1996, Theorem 2.5.2) to conclude in part 2 that

$$\mathbb{G}_n[\zeta_i^*(q; \mathbf{L}, \boldsymbol{\eta})] \Rightarrow \mathbb{G}[\zeta_i^*(q; \mathbf{L}, \boldsymbol{\eta})] \quad \text{in} \quad \ell^\infty(\mathbb{S}^2) \ . \tag{46}$$

Recall $\boldsymbol{\eta} = (\Delta\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\theta}_0)$ and $\widehat{\boldsymbol{\eta}} = (\Delta\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}_0)$. In what follows, we use the explicit notation $(\Delta\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{\theta}_0)$ for the vector of nuisance functions instead of $\boldsymbol{\eta}$, as we analyze the impact on the asymptotic distribution of each of its components.

Part 1: Our goal is to establish Eq. (45). We divide the proof in three claims.

Claim 1: $\sup_{q\in\mathbb{S}^2} \sqrt{n}|\widehat{h}_{\mathcal{E}}(q; \widehat{\mathbf{L}}, (\Delta\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}_0)) - \widehat{h}_{\mathcal{E}}(q; \widehat{\mathbf{L}}, (\Delta\boldsymbol{\theta}, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}_0))| = o_p(1)$. Algebra shows

$$\widehat{h}_{\mathcal{E}}(q; \widehat{\mathbf{L}}, (\Delta\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}_0)) - \widehat{h}_{\mathcal{E}}(q; \widehat{\mathbf{L}}, (\Delta\boldsymbol{\theta}, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}_0)) = \frac{1}{K}\sum_{k=1}^{K}(A_k + B_k),$$

where

$$A_k = \frac{1}{n_k}\sum_{i\in\mathcal{I}_k}\left(\frac{q^\intercal \Delta \mathbf{L}_i Z_i}{\boldsymbol{\pi}(X_i)}\right)\left(\mathbb{1}\{q^\intercal \Delta\widehat{\boldsymbol{\theta}}_k(X_i)>0\} - \mathbb{1}\{q^\intercal \Delta\boldsymbol{\theta}(X_i)>0\}\right) \tag{47}$$

$$B_k = \frac{1}{n_k}\sum_{i\in\mathcal{I}_k}\left(\frac{q^\intercal \Delta \widehat{\mathbf{L}}_i Z_i}{\widehat{\boldsymbol{\pi}}_k(X_i)} - \frac{q^\intercal \Delta \mathbf{L}_i Z_i}{\boldsymbol{\pi}(X_i)}\right)\left(\mathbb{1}\{q^\intercal \Delta\widehat{\boldsymbol{\theta}}_k(X_i)>0\} - \mathbb{1}\{q^\intercal \Delta\boldsymbol{\theta}(X_i)>0\}\right) \tag{48}$$

Lemma B.6 shows that $A_k = o_p(n^{-1/2})$ uniformly on $q \in \mathbb{S}^2$ by relying on Assumptions 4 and 5 and the approach used in the proof of Theorem 4.1 in Liu and Molinari (2026). Using similar ideas, Lemma B.7 shows that $B_k = o_p(n^{-1/2})$ uniformly on $q \in \mathbb{S}^2$. Since K is fixed as $n \to \infty$, we conclude the claim of step 1.

Claim 2: $\sup_{q\in\mathbb{S}^2}\sqrt{n}|\widehat{h}_{\mathcal{E}}(q,\widehat{\mathbf{L}};(\Delta\boldsymbol{\theta},\widehat{\boldsymbol{\pi}},\widehat{\boldsymbol{\theta}}_0)) - n^{-1}\sum_{i=1}^n \zeta_i^*(q;\mathbf{L},(\Delta\boldsymbol{\theta},\widehat{\boldsymbol{\pi}},\widehat{\boldsymbol{\theta}}_0))| = o_p(1)$. Let us denote by $\widehat{\mathcal{M}}_i = \operatorname{diag}\left(\left[1-\frac{\mathbb{1}\{G_i=r\}}{\widehat{\mu}_r}, 1-\frac{\mathbb{1}\{G_i=b\}}{\widehat{\mu}_b}, \frac{\mathbb{1}\{G_i=b\}}{\widehat{\mu}_b}-\frac{\mathbb{1}\{G_i=r\}}{\widehat{\mu}_r}\right]\right)$ the sample analog of $\mathcal{M}$, where μ_g is replaced by its plug-in estimator $\widehat{\mu}_g = n^{-1}\sum_{i=1}^n \mathbb{1}\{G_i=g\}$. Define $\mathcal{M}^r = \operatorname{diag}\left(\left[\frac{1}{\mu_r},0,\frac{1}{\mu_r}\right]\right)$ and $\mathcal{M}^b = \operatorname{diag}\left(\left[0,\frac{1}{\mu_b},-\frac{1}{\mu_b}\right]\right)$, with sample analogs denoted by $\widehat{\mathcal{M}}^r$ and $\widehat{\mathcal{M}}^b$, respectively. Note that $n^{-1}\sum_{i=1}^n \widehat{\mathcal{M}}_i = 0$, which implies

$$\widehat{h}_{\mathcal{E}}(q,\widehat{\mathbf{L}};(\Delta\boldsymbol{\theta},\widehat{\boldsymbol{\pi}},\widehat{\boldsymbol{\theta}}_0)) = \frac{1}{n}\sum_{i=1}^n\left[\zeta_i(q;\widehat{\mathbf{L}},(\boldsymbol{\theta},\widehat{\boldsymbol{\pi}},\widehat{\boldsymbol{\alpha}}^q)) + (\widehat{\mathcal{M}}_i q)^\intercal \nabla_q h_{\mathcal{E}}(q)\right].$$

Therefore, we can write

$$\widehat{h}_{\mathcal{E}}(q,\widehat{\mathbf{L}};(\Delta\boldsymbol{\theta},\widehat{\boldsymbol{\pi}},\widehat{\boldsymbol{\theta}}_0)) - \frac{1}{n}\sum_{i=1}^n \zeta_i^*(q;\mathbf{L},(\Delta\boldsymbol{\theta},\widehat{\boldsymbol{\pi}},\widehat{\boldsymbol{\theta}}_0)) = \sum_{g\in\{r,b\}}\left(\widehat{\mathcal{M}}^g q - \mathcal{M}^g q\right)^\top\left(\frac{1}{K}\sum_{k=1}^K A_k^g + B_k^g\right),$$

where

$$A_k^g = \frac{1}{n_k}\sum_{i\in\mathcal{I}_k}\left\{(\mathbf{L}_0 + (\mathbf{L}_1-\mathbf{L}_0)\mathbb{1}\{q^\intercal\Delta\boldsymbol{\theta}(X_i)>0\})\frac{Z_i\mu_g}{\boldsymbol{\pi}(X_i)} - \nabla_q h_i\right\}\mathbb{1}\{G_i=g\}$$

$$B_k^g = \frac{1}{n_k}\sum_{i\in\mathcal{I}_k}\left(\frac{Z_i\mu_g}{\widehat{\boldsymbol{\pi}}_k(X_i)} - \frac{Z_i\mu_g}{\boldsymbol{\pi}(X_i)}\right)(\mathbf{L}_0 + (\mathbf{L}_1-\mathbf{L}_0)\mathbb{1}\{q^\intercal\Delta\boldsymbol{\theta}(X_i)>0\})\,\mathbb{1}\{G_i=g\}$$

$$\nabla_q h_i = \operatorname{diag}\left([\mathbb{1}\{G_i=r\},\mathbb{1}\{G_i=b\},1]\right)\nabla_q h_{\mathcal{E}}(q)\,.$$

We conclude that $A_k^g = O_p(n^{-1/2})$ uniformly on $q\in\mathbb{S}^2$ by Assumptions 1 and 3, Eq. (9) adjusted with inverse propensity weighting similarly to Lemma 3.1, and the Central Limit

Theorem. We also conclude that $B_k^g = o_p(1)$ uniformly on $q \in \mathbb{S}^2$ by Cauchy-Schwartz, and Assumptions 1, 4, and 5. Finally, we conclude $\mathcal{M}^g q - \widehat{\mathcal{M}}^g q = O_p(n^{-1/2})$ uniformly on $q \in \mathbb{S}^2$ by the Central Limit Theorem and Assumption 1.

Claim 3: $\sup_{q\in\mathbb{S}^2} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\zeta_i^*(q; \mathbf{L}, (\Delta\boldsymbol{\theta}, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}_0)) - h_{\mathcal{E}}(q) \right) - \mathbb{G}_n[\zeta_i^*(q; \mathbf{L}, \boldsymbol{\eta})] \right| = o_p(1)$.

We write

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \left(\zeta_i^*(q; \mathbf{L}, (\Delta\boldsymbol{\theta}, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}_0)) - h_{\mathcal{E}}(q) \right) - \mathbb{G}_n[\zeta_i^*(q; \mathbf{L}, \boldsymbol{\eta})] = \frac{1}{\sqrt{K}} \sum_{k=1}^K (A_k + B_k + C_k)$$

where

$$A_k = \frac{1}{\sqrt{n_k}} \sum_{i\in\mathcal{I}_k} \left(\frac{Z_i}{\widehat{\boldsymbol{\pi}}_k(X_i)} - \frac{Z_i}{\boldsymbol{\pi}(X_i)} \right) \left(q^\intercal \mathbf{L}_{0,i} + (q^\intercal \Delta\mathbf{L}_i)\, \mathbb{1}\{q^\intercal \Delta\boldsymbol{\theta}(X_i) > 0\} - \boldsymbol{\alpha}^h(q; X_i) \right),$$

$$B_k = \frac{1}{\sqrt{n_k}} \sum_{i\in\mathcal{I}_k} \left(\widehat{\boldsymbol{\alpha}}_k^h(q; X) - \boldsymbol{\alpha}^h(q; X) \right) \left(1 - \frac{Z_i}{\boldsymbol{\pi}(X_i)} \right),$$

$$C_k = \frac{1}{\sqrt{n_k}} \sum_{i\in\mathcal{I}_k} \left(\widehat{\boldsymbol{\alpha}}_k^h(q; X) - \boldsymbol{\alpha}^h(q; X) \right) \left(\frac{Z_i}{\boldsymbol{\pi}(X_i)} - \frac{Z_i}{\widehat{\boldsymbol{\pi}}_k(X_i)} \right).$$

Recall that $\widehat{\boldsymbol{\eta}}_k = (\Delta\widehat{\boldsymbol{\theta}}_k, \widehat{\boldsymbol{\pi}}_k, \boldsymbol{\theta}_{0,k})$ was estimated using all the data except the ones with indices in $\mathcal{I}_k$, and $\mathcal{F}_k$ is the σ-algebra generated by all the data except the ones with indices on $\mathcal{I}_k$. Assumptions 4 and 5, the definition of $\boldsymbol{\alpha}^h(q; X)$ in Eq. (27), and algebra show $\mathbb{E}[A_k|\mathcal{F}_k] = 0$ and $\sup_{q\in\mathbb{S}^2} \mathbb{E}[A_k^2|\mathcal{F}_k] = o(1)$. Similarly, we can show $\mathbb{E}[B_k|\mathcal{F}_k] = 0$ and $\sup_{q\in\mathbb{S}^2} \mathbb{E}[B_k^2|\mathcal{F}_k] = o(1)$. By Chernozhukov et al. (2018, Lemma 6.1), we conclude A_k and B_k both are $o_p(1)$ uniformly on $q \in \mathbb{S}^2$. Finally, we conclude that $C_k = o_p(1)$ uniformly on $q \in \mathbb{S}^2$ by Cauchy-Schwartz, Assumptions 4 and 5, and Lemma B.8.

The triangle inequality and Claims 1, 2, and 3, complete the proof of part 1.

Part 2: Our goal is to establish Eq. (46). Consider the random vector

$$W_i = (\mathbf{L}_{0,i}, \Delta\mathbf{L}_i, Z_i, \boldsymbol{\pi}(X_i), \Delta\boldsymbol{\theta}(X_i), \boldsymbol{\theta}_0(X_i)) \in \mathcal{W}$$

and the class of functions $\mathcal{Q} \equiv \{f : \mathcal{W} \to \mathbb{R} : f(W_i) = \zeta_i(q; \mathbf{L}, \boldsymbol{\eta}) \text{ for some } q \in \mathbb{S}^2\}$. Here, we are using the true value of the nuisance function $\boldsymbol{\eta}$ and treating the values $\boldsymbol{\eta}(X_i)$ as a random vector. In order to use van der Vaart and Wellner (1996, Theorem 2.5.2) to complete the proof of part 2, we need to verify that (i) an envelope function of the class $\mathcal{Q}$ has finite

second-moment, and (ii) the uniform entropy bound in van der Vaart and Wellner (1996, Eq. (2.5.1)) holds. To verify (i), we note that the minimum envelope $\sup_{q\in\mathbb{S}^2}|\zeta_i(q;\mathbf{L},\boldsymbol{\eta})|$ has finite second-moment due to Assumptions 1 and 4. To verify (ii), we rely on Andrews (1994, Theorems 2 and 3) that states that the class of functions of type I and their mixing (by addition and product operation) verify the uniform entropy bound. We conclude by noting that $\mathcal{Q}$ is of type I, since any function in $\mathcal{Q}$ can be written as the sum of (i) linear functions in q, (ii) indicators of linear functions in q, and (iii) product of (i) and (ii). *Q.E.D.*

A.3. *Proofs for Section 4*

Let $[\cdot]_+ = \max\{\cdot, 0\}$. Define $\phi : \ell^\infty(\mathbb{S}^2)\times\mathbb{R}^3 \to \mathbb{R}$ as

$$\phi(f,\varepsilon) \equiv \left[\max_{q\in\mathbb{S}^2}(q^\intercal\varepsilon - f(q))\right]_+ + \left[\min_{q\in\widetilde{\mathbb{S}}^2}(q_1\varepsilon_1 + q_2\varepsilon_2 + |q_3\varepsilon_3| + f(-q))\right]_+, \tag{49}$$

where $f\in\ell^\infty(\mathbb{S}^2)$ and $\varepsilon\in\mathbb{R}^3$. Note that $\varphi_n^{\text{LDA}} = \sqrt{n}\left(\phi(\widehat{\boldsymbol{\beta}}) - \phi(\boldsymbol{\beta})\right)$.

PROOF OF THEOREM 4.1: The proof has two parts. In the first part, we use Fang and Santos (2019, Theorems 2.1 and 3.2) to show that the Bayesian bootstrap is consistent. In the second part, we use Romano and Shaikh (2012, Lemma A.1) to prove that the test φ_n^{LDA} has asymptotically correct size.

$$F_n(x) = \mathbb{P}\left(\sqrt{n}\left(\phi(\widehat{\boldsymbol{\beta}}) - \phi(\boldsymbol{\beta})\right)\le x\right)$$
$$\widetilde{F}_n(x) = \mathbb{P}\left(\widehat{\phi}'_{\boldsymbol{\beta}}\left(\sqrt{n}\{\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\}\right)\le x | \{(Y_i,G_i,X_i,Z_i)\}_{i=1}^n\right)$$

Part 1: By Kaido (2016, Lemma D.3) and standard arguments, it follows that ϕ defined in (49) is Hadamard directionally differentiable at $\boldsymbol{\beta}$ tangentially to $\mathbb{D}_0 = C(\mathbb{S}^2)\times\mathbb{R}^3 \subset \mathbb{D} = \ell^\infty(\mathbb{S}^2)\times\mathbb{R}^3$, where $C(\mathbb{S}^2)$ is the class of continuous functions on $\mathbb{S}^2$. This guarantees Assumption 1 in Fang and Santos (2019). By Cramer-Wold and Lemma B.10, we have $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \Rightarrow \mathbb{G}_{\boldsymbol{\beta}}$. Let $K(q,q')$ be the covariance matrix of $\mathbb{G}_{\boldsymbol{\beta}}$ for $q,q'\in\mathbb{S}^2$. It can be show that $K(q,q')$ is continuous in both $q,q'\in\mathbb{S}^2$, which implies that $\mathbb{G}_{\boldsymbol{\beta}}\in\mathbb{D}_0$

with probability 1 by Adler and Taylor (2007, Lemma 1.3.1). This guarantees Assumption 2 in Fang and Santos (2019). By Fang and Santos (2019, Theorems 2.1), we conclude $\sqrt{n}\left(\phi(\widehat{\boldsymbol{\beta}}) - \phi(\boldsymbol{\beta})\right) \Rightarrow \phi'_{\boldsymbol{\beta}}(\mathbb{G}_{\boldsymbol{\beta}})$.

Define $F_\infty(x) = \mathbb{P}\left(\phi'_{\boldsymbol{\beta}}(\mathbb{G}_{\boldsymbol{\beta}}) \leq x\right)$. To prove the consistency of the bootstrap, it is sufficient to show $\mathbb{P}\left(\sup_{x\in\mathbb{R}} \widetilde{F}_n(x) - F_\infty(x) > \epsilon\right) = o(1)$ for any $\epsilon > 0$, which is equivalent to show that

$$\sup_{f\in\mathcal{BL}_1} \left| \mathbb{E}\left[f\left(\widehat{\phi}'_{\boldsymbol{\beta}}\left(\sqrt{n}\{\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\}\right)\right) \mid \{(Y_i, G_i, X_i, Z_i)\}_{i=1}^n \right] - \mathbb{E}\left[f\left(\phi'_{\boldsymbol{\beta}}(\mathbb{G}_{\boldsymbol{\beta}})\right)\right] \right| = o_p(1), \quad (50)$$

where $\mathcal{BL}_1$ is the set of 1-Lipschitz functions $f : \mathbb{R} \to \mathbb{R}$ such that $|f|_\infty \leq 1$.

It is sufficient to verify Assumptions 3 and 4 in Fang and Santos (2019) to conclude (50) by Fang and Santos (2019, Theorems 3.2). Note that Assumption 3 part (i), (iii), and (iv) hold by construction. By Hong and Li (2018, Theorem 3.1), we conclude $\widehat{\phi}'_{\boldsymbol{\beta}}$ defined in Procedure 1 verifies Assumption 4 in Fang and Santos (2019). Therefore, it is sufficient to verify Assumption 3 part (ii), that is,

$$\sup_{f\in\mathcal{BL}_1} \left| \mathbb{E}\left[f\left(\sqrt{n}\left(\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\right)\right) \mid \{(Y_i, G_i, X_i, Z_i)\}_{i=1}^n \right] - \mathbb{E}\left[f\left(\mathbb{G}_{\boldsymbol{\beta}}\right)\right] \right| = o(1)\,. \quad (51)$$

Define $\widehat{\boldsymbol{\beta}}^* = \left(\frac{1}{n}\sum_{i=1}^n \left(\frac{W_i}{\overline{W}}\right) \zeta_i^*(q; \mathbf{L}, \boldsymbol{\eta}), \frac{1}{n}\sum_{i=1}^n \left(\frac{W_i}{\overline{W}}\right) \xi_i^*(a; \mathbf{L}, \boldsymbol{\eta}) \right)^\intercal$. By Lemma B.9 and since $f \in \mathcal{BL}_1$, we conclude that Eq. (51) holds if the following equation holds

$$\sup_{f\in\mathcal{BL}_1} \left| \mathbb{E}\left[f\left(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}^* - \widehat{\boldsymbol{\beta}}\right)\right) \mid \{(Y_i, G_i, X_i, Z_i)\}_{i=1}^n \right] - \mathbb{E}\left[f\left(\mathbb{G}_{\boldsymbol{\beta}}\right)\right] \right| = o(1)\,.$$

Finally, the previous equation holds by definition of $\widehat{\boldsymbol{\beta}}^*$, $\widehat{\boldsymbol{\beta}}$, and by van der Vaart and Wellner (1996, Theorem 3.6.13). This completes the proof of the bootstrap consistency.

Part 2: For an arbitrarily small $\kappa > 0$, we have $\mathbb{P}\left(\sup_{x\in\mathbb{R}} \widetilde{F}_n(x) - F_\infty(x) > \kappa\right) = o(1)$. Let $c_{1-\alpha}$ be $(1-\alpha)$-quantile of $\phi'_{\boldsymbol{\beta}}(\mathbb{G}_{\boldsymbol{\beta}})$, i.e., $F_\infty(c_{1-\alpha}) \geq 1-\alpha$. Recall that $\hat{c}_{1-\alpha+\kappa}$ is the $(1-\alpha+\kappa)-$quantile of the limit distribution of $\widehat{\phi}'_{\boldsymbol{\beta}}\left(\sqrt{n}\{\widetilde{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}\}\right)$. The following derivation completes the proof of part 2:

$$\mathbb{P}(T_n^{\text{LDA}} > \hat{c}_{1-\alpha+\kappa} + \kappa) \overset{(1)}{=} \mathbb{P}(T_n^{\text{LDA}} > c_{1-\alpha} + \kappa) + o(1)$$

$$\overset{(2)}{\leq} \mathbb{P}(T_n^{\text{LDA}} \geq c_{1-\alpha} + \kappa/2) + o(1)$$

$$\overset{(3)}{\leq} \mathbb{P}(\phi'_{\boldsymbol{\beta}}(\mathbb{G}_{\boldsymbol{\beta}}) \geq c_{1-\alpha} + \kappa/2) + o(1)$$

$$\overset{(4)}{\leq} \mathbb{P}(\phi'_{\boldsymbol{\beta}}(\mathbb{G}_{\boldsymbol{\beta}}) > c_{1-\alpha}) + o(1)$$

$$\overset{(5)}{\leq} \alpha + o(1)\ ,$$

where (1) holds since $\mathbb{P}(\hat{c}_{1-\alpha+\kappa} \geq c_{1-\alpha}) = 1 - o(1)$ by Romano and Shaikh (2012, Lemma A.1), (2) and (4) hold since $\kappa > 0$, (3) holds since $T_n^{\text{LDA}} \xrightarrow{d} \phi'_{\boldsymbol{\beta}}(\mathbb{G}_{\boldsymbol{\beta}})$, and (5) holds by the definition of $c_{1-\alpha}$. *Q.E.D.*

APPENDIX B: AUXILIARY RESULTS

B.1. *Lemmas and Auxiliary Results for Section 2*

LEMMA B.1—Support function of $\mathcal{C}(\varepsilon^*)$: For $\varepsilon^* = (\varepsilon_1^*, \varepsilon_2^*, \varepsilon_3^*) \in \mathcal{E}$ and $q \in \mathbb{R}^3$, the support function of the closed convex set $\mathcal{C}(\varepsilon^*)$ in Eq. (10) is as given in Eq. (11).

PROOF: We have:

$$h_{\mathcal{C}(\varepsilon^*)}(q) = \sup_{\varepsilon \in \mathcal{C}(\varepsilon^*)} q^\intercal \varepsilon = \sup_{\substack{\varepsilon_1 \leq \varepsilon_1^*, \varepsilon_2 \leq \varepsilon_2^* \\ |\varepsilon_3| \leq |\varepsilon_3^*|}} (q_1\varepsilon_1 + q_2\varepsilon_2 + q_3\varepsilon_3).$$

If $q_1 < 0$, then $\sup_{\varepsilon_1 \leq \varepsilon_1^*} q_1\varepsilon_1 = +\infty$ (taking $\varepsilon_1 \to -\infty$). Similarly if $q_2 < 0$. For $q_1, q_2 \geq 0$, the supremum over $\varepsilon_j \leq \varepsilon_j^*$ gives $q_j\varepsilon_j^*$, $j = 1, 2$. For the third coordinate:

$$\sup_{|\varepsilon_3| \leq |\varepsilon_3^*|} q_3\varepsilon_3 = \begin{cases} q_3|\varepsilon_3^*| & \text{if } q_3 \geq 0, \\ -q_3|\varepsilon_3^*| = |q_3||\varepsilon_3^*| & \text{if } q_3 < 0. \end{cases}$$

Thus $h_{\mathcal{C}(\varepsilon^*)}(q) = q_1\varepsilon_1^* + q_2\varepsilon_2^* + |q_3||\varepsilon_3^*|$ for $q_1, q_2 \geq 0$. *Q.E.D.*

LEMMA B.2: Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be the probability space on which (Y, G, X) are defined. Assume that $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is a standard Borel space. Then: (i) $\mathcal{E}$ is convex. (ii) Under Assumption 1, $\mathcal{E}$ is compact. (iii) Under Assumptions 1-2, $\mathcal{E}$ has a non-empty interior in $\mathbb{R}^3$, and hence $\dim(\mathcal{E}) = 3$.

PROOF: Recall that $\mathcal{A} \equiv \{a : \mathcal{X} \to [0,1] \text{ Borel-measurable}\}$.

Part (i). Each coordinate map $a \mapsto e_g^\iota(a)$ is affine in a by (3). Since $\mathcal{A}$ is convex and $(e_r^A(a), e_b^A(a), e_r^F(a) - e_b^F(a))$ is an affine image of a, its range $\mathcal{E}$ is convex.

Part (ii). Define

$$K \equiv \{u \in L^\infty(\Omega, \mathfrak{F}, \mathbb{P}) : u \text{ is } \sigma(X)\text{-measurable and } 0 \leq u \leq 1 \text{ a.s.}\}.$$

As $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is standard Borel, the Doob–Dynkin lemma (Rao and Swift, 2006, Prop. 3, Ch. 1) yields $K = \{a(X) : a \in \mathcal{A}\}$ up to a.s. equality. Hence, $\mathcal{E} = \mathbb{E}[\boldsymbol{\theta}_0(X)] + T(K)$, for

$$T(u) \equiv \mathbb{E}[u\Delta\boldsymbol{\theta}(X)] \in \mathbb{R}^3. \tag{52}$$

We first show that K is weak-* compact in L^∞, where L^∞ is equipped with the topology $\sigma(L^\infty, L^1)$. Since $0 \leq u \leq 1$ a.s. for every $u \in K$, we have $\|u\|_\infty \leq 1$, so K is contained in the closed unit ball of L^∞. Since $L^\infty = (L^1)^*$, the Banach-Alaoglu theorem yields that the closed unit ball of L^∞ is compact for the weak-* topology $\sigma(L^\infty, L^1)$.

To prove weak-* compactness of K, it suffices to show that K is weak-* closed, since K is contained in the weak-* compact unit ball of L^∞. Accordingly, let $u_\alpha \in K$ and suppose $u_\alpha \to u$ weak-* in L^∞. We show that $u \in K$. First, u is $\sigma(X)$-measurable. Indeed, for every $g \in L^1(\Omega, \mathfrak{F}, \mathbb{P})$, $\mathbb{E}\left[u_\alpha\big(g - \mathbb{E}[g|\sigma(X)]\big)\right] = 0$, because u_α is $\sigma(X)$-measurable. Passing to the limit using weak-* convergence gives $\mathbb{E}\left[u\left(g - \mathbb{E}[g|\sigma(X)]\right)\right] = 0$, for all $g \in L^1$. Taking $g = \mathbb{1}_A$ for $A \in \mathfrak{F}$, we obtain

$$\int_A u d\mathbb{P} = \int_A \mathbb{E}[u|\sigma(X)] d\mathbb{P}, \quad \forall A \in \mathfrak{F},$$

hence $u = \mathbb{E}[u|\sigma(X)]$ a.s. Therefore u is $\sigma(X)$-measurable.

Next, the bounds $0 \leq u \leq 1$ a.s. are preserved. To show $u \geq 0$ a.s., suppose otherwise. Then $\mathbb{P}(u < 0) > 0$, so for some $\varepsilon > 0$, $A_\varepsilon \equiv \{u \leq -\varepsilon\}$ has positive probability. Taking $g = \mathbb{1}_{A_\varepsilon} \in L^1$, weak-* convergence yields $\mathbb{E}[u_\alpha g] \to \mathbb{E}[ug]$. Since $u_\alpha \geq 0$ a.s., we have $\mathbb{E}[u_\alpha g] \geq 0$ for all α, whereas $\mathbb{E}[ug] = \mathbb{E}[u\mathbb{1}_{A_\varepsilon}] \leq -\varepsilon\,\mathbb{P}(A_\varepsilon) < 0$, a contradiction. Thus $u \geq 0$ a.s. Applying the same argument to $1 - u_\alpha$ shows $u \leq 1$ a.s. Therefore $u \in K$, so K is weak-* closed. Hence K is weak-* compact.

Now consider the map $T : K \to \mathbb{R}^3$ in (52). This map is well-defined because $u \in L^\infty$ with $|u| \leq 1$ a.s. and each component of $\Delta\boldsymbol{\theta}(X)$ is in L^1 by Assumption 1, so each component of $u\Delta\boldsymbol{\theta}(X)$ is in L^1. Moreover, for each component $j = 1, 2, 3$, the map from u to the component of $\mathbb{E}[u\Delta\boldsymbol{\theta}(X)]$ is weak-* continuous on L^∞, since each component of $\Delta\boldsymbol{\theta}(X)$ is in L^1. Therefore $T : K \to \mathbb{R}^3$ is continuous when K is endowed with the weak-* topology and $\mathbb{R}^3$ with its usual Euclidean topology. Since K is weak-* compact and T is continuous, $T(K)$ is compact in $\mathbb{R}^3$. Therefore, $\mathcal{E} = \mathbb{E}[\boldsymbol{\theta}_0(X)] + T(K)$ is compact.

Part (iii). Let $\mathcal{A}_{\text{disp}} \equiv \{\varepsilon : \mathcal{X} \to \mathbb{R} \text{ measurable} : \ |\varepsilon(X)| \leq 1/2 \ P_X\text{-a.s.}\} = \{\varepsilon : \mathcal{X} \to \mathbb{R} \text{ measurable} : \ \|\varepsilon\|_\infty \leq 1/2\}$ where $\|\varepsilon\|_\infty \equiv \operatorname{ess\,sup}_{x \sim P_X} |\varepsilon(x)|$. For any algorithm $a \in \mathcal{A}$, let $\varepsilon(X) \equiv a(X) - \frac{1}{2}$ be the corresponding displacement in $\mathcal{A}_{\text{disp}}$. Let $e^0 = \mathbb{E}\big[\boldsymbol{\theta}_0(X)\big] + \frac{1}{2}\mathbb{E}[\Delta\boldsymbol{\theta}(X)]$. Then for any $a \in \mathcal{A}$ we have $e(a) - e^0 = \mathbb{E}[\varepsilon(X)\Delta\boldsymbol{\theta}(X)]$. Hence, we can write $\mathcal{E} = e^0 + \mathcal{E}^0$, where $\mathcal{E}^0$ is the set of feasible displacements:

$$\mathcal{E}^0 \equiv \left\{\mathbb{E}[\varepsilon(X)\Delta\boldsymbol{\theta}(X)],\ \varepsilon : \mathcal{X} \to [-\tfrac{1}{2}, \tfrac{1}{2}] \text{ measurable}\right\},$$

and the equality holds because every measurable $a : \mathcal{X} \to [0, 1]$ corresponds to a unique $\varepsilon \in [-\frac{1}{2}, \frac{1}{2}]$ and vice versa. We next show that $\mathcal{E}^0$ contains a Euclidean ball around 0, which implies $\mathcal{E} = e^0 + \mathcal{E}^0$ contains a ball around e^0. Fix $q \in \mathbb{R}^3$ with $\|q\| = 1$. Consider the support function of $\mathcal{E}^0$ in direction q:

$$h_{\mathcal{E}^0}(q) \equiv \sup_{e \in \mathcal{E}^0} q^\intercal e = \sup_{\|\varepsilon\|_\infty \leq 1/2} \mathbb{E}\left[\varepsilon(X) q^\intercal \Delta\boldsymbol{\theta}(X)\right].$$

Pointwise maximization yields the maximizer $\varepsilon_q(X) = \frac{1}{2}\operatorname{sign}\big(q^\intercal\Delta\boldsymbol{\theta}(X)\big)$ (with any value in $[-\frac{1}{2}, \frac{1}{2}]$ on the null set where $q^\intercal\Delta\boldsymbol{\theta}(X) = 0$), so

$$h_{\mathcal{E}^0}(q) = \frac{1}{2}\mathbb{E}\left[|q^\intercal\Delta\boldsymbol{\theta}(X)|\right]. \tag{53}$$

By Assumption 2, for all $\delta > 0$, $\mathbb{P}(|q^\intercal\Delta\boldsymbol{\theta}(X)| \leq \delta) \leq C\delta^m$, uniformly in $q \in \mathbb{S}^2 \equiv \{q \in \mathbb{R}^3 : \|q\| = 1\}$. Choose $\delta_0 \equiv (1/(2C))^{1/m} > 0$. Then for every $q \in \mathbb{S}^2$, $\mathbb{P}(|q^\intercal\Delta\boldsymbol{\theta}(X)| > \delta_0) \geq 1 - C\delta_0^m = \frac{1}{2}$, and hence $\mathbb{E}\left[|q^\intercal\Delta\boldsymbol{\theta}(X)|\right] \geq \delta_0\mathbb{P}(|q^\intercal\Delta\boldsymbol{\theta}(X)| > \delta_0) \geq \frac{\delta_0}{2}$. Combining

with (53) yields the uniform bound

$$h_{\mathcal{E}^0}(q) \geq \tfrac{\delta_0}{4} \qquad \forall q \in \mathbb{S}^2. \tag{54}$$

Hence, (54) implies that $\mathcal{E}^0$ contains the Euclidean ball $\mathbb{B}(0,r)$ with radius $r = \delta_0/4$ (see, e.g., Schneider, 1993, Section 1.7). Therefore, $e^0 + \mathbb{B}(0,r) \subseteq \mathcal{E}$ so $\mathcal{E}$ has a nonempty interior in $\mathbb{R}^3$, and hence $\dim(\mathcal{E}) = 3$. *Q.E.D.*

B.1.1. *Lemmas and Auxiliary Results for the Geometry of $\mathcal{F}$*

We distinguish three cases used to characterize the geometric properties of the FA-frontier.

CASE 1: The feasible set $\mathcal{E}$ lies entirely on one side of the plane $\varepsilon_3 = 0$.

THEOREM B.1—FA-frontier in Case 1: *Suppose Assumptions 1 and 2 hold, and $\mathcal{E}$ is in Case 1. Without loss of generality, assume $\mathcal{E} \subset \{\varepsilon \in \mathbb{R}^3 : \varepsilon_3 \geq 0\}$.*[10] *For $\varepsilon^* \in \mathcal{E}$, define*

$$\mathcal{C}_+(\varepsilon^*) \equiv \{\varepsilon \in \mathbb{R}^3 : \varepsilon_1 \leq \varepsilon_1^*,\ \varepsilon_2 \leq \varepsilon_2^*,\ \varepsilon_3 \leq \varepsilon_3^*\},$$

and let $\widetilde{\mathbb{S}}^2_+ \equiv \{q \in \mathbb{S}^2 : q_1 \geq 0,\ q_2 \geq 0,\ q_3 \geq 0\}$. Then the following are equivalent:

1. *$\mathcal{E} \cap \mathcal{C}_+(\varepsilon^*) = \{\varepsilon^*\}$.*
2. *There exists $q \in \widetilde{\mathbb{S}}^2_+$ such that $h_{\mathcal{C}_+(\varepsilon^*)}(q) = -h_{\mathcal{E}}(-q)$.*
3. *There exists $q \in \widetilde{\mathbb{S}}^2_+$ such that $\varepsilon^* = \arg\min_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon$.*

Consequently, in Case 1 the FA-frontier satisfies

$$\mathcal{F} = \{\varepsilon^* \in \mathcal{E} : \mathcal{E} \cap \mathcal{C}_+(\varepsilon^*) = \{\varepsilon^*\}\} = \bigcup_{q \in \widetilde{\mathbb{S}}^2_+} \arg\min_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon,$$

Since $\mathcal{E}$ is strictly convex, each set $\arg\min_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon$ is a singleton.

[10] This is without loss of generality because Definition 5 depends on the third coordinate only through $|\varepsilon_3|$, hence the reflection $(\varepsilon_1, \varepsilon_2, \varepsilon_3) \mapsto (\varepsilon_1, \varepsilon_2, -\varepsilon_3)$ leaves the FA-dominance relation and the FA-frontier unchanged.

PROOF: As $\mathcal{E}$ is in Case 1, every $\varepsilon \in \mathcal{E}$ satisfies $\varepsilon_3 \geq 0$. Therefore, for $\varepsilon, \varepsilon' \in \mathcal{E}$, $|\varepsilon'_3| \leq |\varepsilon_3| \iff \varepsilon'_3 \leq \varepsilon_3$. Hence, $\varepsilon' \succ^{FA} \varepsilon^* \iff \varepsilon' \in \mathcal{E} \cap \mathcal{C}_+(\varepsilon^*) \setminus \{\varepsilon^*\}$, and so $\varepsilon^* \in \mathcal{F} \iff \mathcal{E} \cap \mathcal{C}_+(\varepsilon^*) = \{\varepsilon^*\}$. We note that the support function of $\mathcal{C}_+(\varepsilon^*)$ satisfies $h_{\mathcal{C}_+(\varepsilon^*)}(q) = q^\intercal \varepsilon^*$ for $q_1, q_2, q_3 \geq 0$ and $h_{\mathcal{C}_+(\varepsilon^*)}(q) = +\infty$ otherwise, because $\mathcal{C}_+(\varepsilon^*)$ is unbounded in the negative coordinate directions.

We begin by showing (1) $\Rightarrow$ (2). Assume $\mathcal{E} \cap \mathcal{C}_+(\varepsilon^*) = \{\varepsilon^*\}$. By Lemma B.2, $\mathcal{E}$ is compact, convex, and has a nonempty interior. Hence, by the same common supporting hyperplane argument used in Step 3 of the proof of Theorem 2.2, there exist $q \neq 0$ and $\alpha \in \mathbb{R}$ such that $q^\intercal \varepsilon \geq \alpha \geq q^\intercal c$, $\forall \varepsilon \in \mathcal{E}$, $\forall c \in \mathcal{C}_+(\varepsilon^*)$. Since $\varepsilon^* \in \mathcal{E} \cap \mathcal{C}_+(\varepsilon^*)$, necessarily $\alpha = q^\intercal \varepsilon^*$. Moreover, $\inf_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon \geq \alpha \geq \sup_{c \in \mathcal{C}_+(\varepsilon^*)} q^\intercal c$, while $\inf_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon \leq q^\intercal \varepsilon^* = \alpha \leq \sup_{c \in \mathcal{C}_+(\varepsilon^*)} q^\intercal c$. Therefore, $-h_{\mathcal{E}}(-q) = h_{\mathcal{C}_+(\varepsilon^*)}(q)$. Because $h_{\mathcal{C}_+(\varepsilon^*)}(q) < +\infty$, this implies $q_1, q_2, q_3 \geq 0$. By positive homogeneity, after replacing q by $q/\|q\|$, we may assume $q \in \widetilde{\mathbb{S}}^2_+$, proving (2).

We show that (2) $\Rightarrow$ (3). Assume there exists $q \in \widetilde{\mathbb{S}}^2_+$ such that $h_{\mathcal{C}_+(\varepsilon^*)}(q) = -h_{\mathcal{E}}(-q)$. Then $\sup_{c \in \mathcal{C}_+(\varepsilon^*)} q^\intercal c = \inf_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon$. Because $\varepsilon^* \in \mathcal{C}_+(\varepsilon^*)$, we have

$$q^\intercal \varepsilon^* \leq \sup_{c \in \mathcal{C}_+(\varepsilon^*)} q^\intercal c = \inf_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon \leq q^\intercal \varepsilon^*.$$

Hence all inequalities are equalities, and thus $q^\intercal \varepsilon^* = \inf_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon$. Therefore, $\varepsilon^* \in \arg\min_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon$. Since $\mathcal{E}$ is strictly convex, every support set is a singleton, so $\varepsilon^* = \arg\min_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon$, proving (2) $\Rightarrow$ (3).

We conclude by showing (3) $\Rightarrow$ (1). Assume there exists $q \in \widetilde{\mathbb{S}}^2_+$ such that $\varepsilon^* = \arg\min_{\varepsilon \in \mathcal{E}} q^\intercal \varepsilon$. Let $\varepsilon \in \mathcal{E} \cap \mathcal{C}_+(\varepsilon^*)$. Since $\varepsilon \in \mathcal{C}_+(\varepsilon^*)$, we have $\varepsilon_i \leq \varepsilon^*_i$, $i = 1, 2, 3$. Because $q_i \geq 0$ for each i, $q^\intercal \varepsilon \leq q^\intercal \varepsilon^*$. Since ε^* minimizes $q^\intercal \varepsilon$ over $\mathcal{E}$, we also have $q^\intercal \varepsilon \geq q^\intercal \varepsilon^*$ and hence $q^\intercal \varepsilon = q^\intercal \varepsilon^*$. Therefore ε belongs to the support set of $\mathcal{E}$ in direction $-q$. Strict convexity of $\mathcal{E}$ implies that this support set is a singleton, namely $\{\varepsilon^*\}$. Thus $\varepsilon = \varepsilon^*$, proving (3) $\Rightarrow$ (1).

Finally, the displayed characterization of $\mathcal{F}$ follows immediately from the equivalence just proved. *Q.E.D.*

COROLLARY B.1—Boundary edge images for the FA-frontier in Case 1: *Under the assumptions of Theorem B.1, define the support-point map*

$$\Psi : \widetilde{\mathbb{S}}^2_+ \to \mathcal{F}, \qquad \Psi(q) \equiv \arg\min_{\varepsilon\in\mathcal{E}} q^\intercal \varepsilon. \tag{55}$$

Then Ψ is well-defined and continuous. Moreover, the image under Ψ of the three boundary edges of $\widetilde{\mathbb{S}}^2_+$ is given by the three paths

$$RB(\alpha) \equiv \arg\min_{\varepsilon\in\mathcal{E}} \big[\alpha\varepsilon_1 + (1-\alpha)\varepsilon_2\big], \qquad \alpha \in [0,1] \tag{56}$$

$$RF(\alpha) \equiv \arg\min_{\varepsilon\in\mathcal{E}} \big[\alpha\varepsilon_1 + (1-\alpha)\varepsilon_3\big], \qquad \alpha \in [0,1], \tag{57}$$

$$BF(\alpha) \equiv \arg\min_{\varepsilon\in\mathcal{E}} \big[\alpha\varepsilon_2 + (1-\alpha)\varepsilon_3\big], \qquad \alpha \in [0,1]. \tag{58}$$

If, in addition, for every $\varepsilon \in \mathcal{F}$ the set $N^+_{\mathcal{E}}(\varepsilon) \equiv \{q \in \widetilde{\mathbb{S}}^2_+ : \varepsilon = \arg\min_{\eta\in\mathcal{E}} q^\intercal\eta\}$ is a singleton, then Ψ is injective and hence a homeomorphism from $\widetilde{\mathbb{S}}^2_+$ onto $\mathcal{F}$. In particular, $RB([0,1]) \equiv \{RB(\alpha) : \alpha \in [0,1]\} \subsetneq \mathcal{F}$, and therefore $\mathcal{F} \neq RB([0,1])$.

PROOF: For each $q \in \widetilde{\mathbb{S}}^2_+$, strict convexity of $\mathcal{E}$ (Theorem 2.1) implies that $\arg\min_{\varepsilon\in\mathcal{E}} q^\intercal\varepsilon$ is a singleton, and Theorem B.1 implies that this minimizer belongs to $\mathcal{F}$. Hence, Ψ is well-defined. Continuity follows from Berge's maximum theorem: $\mathcal{E}$ is compact by Lemma B.2, the objective $(q,\varepsilon) \mapsto q^\intercal\varepsilon$ is continuous, and the minimizer is unique for every $q \in \widetilde{\mathbb{S}}^2_+$.

We next identify the images of the three boundary edges of $\widetilde{\mathbb{S}}^2_+$. Consider first the edge

$$\{q \in \widetilde{\mathbb{S}}^2_+ : q_3 = 0\}.$$

Every point on this edge has the form

$$q = \tfrac{(\alpha,1-\alpha,0)}{\|(\alpha,1-\alpha,0)\|} \qquad \text{for some } \alpha \in [0,1],$$

and conversely every $\alpha \in [0,1]$ determines a point on this edge after normalization. Since multiplying a normal vector by a positive scalar does not change the $\arg\min$, we obtain

$$\Psi(q) = \arg\min_{\varepsilon\in\mathcal{E}} [\alpha\varepsilon_1 + (1-\alpha)\varepsilon_2] = RB(\alpha).$$

Therefore,

$$\Psi\left(\{q \in \widetilde{\mathbb{S}}^2_+ : q_3 = 0\}\right) = RB([0,1]).$$

Repeating the same argument to the other two edges, we obtain

$$\Psi\left(\{q \in \widetilde{\mathbb{S}}^2_+ : q_2 = 0\}\right) = RF([0,1]),$$
$$\Psi\left(\{q \in \widetilde{\mathbb{S}}^2_+ : q_1 = 0\}\right) = BF([0,1]).$$

Now assume that $N^+_{\mathcal{E}}(\varepsilon)$ is a singleton for every $\varepsilon \in \mathcal{F}$. If $\Psi(q) = \Psi(q') = \varepsilon$, then $q, q' \in N^+_{\mathcal{E}}(\varepsilon)$, hence $q = q'$. Therefore Ψ is injective. By Theorem B.1, $\mathcal{F} = \Psi(\widetilde{\mathbb{S}}^2_+)$, so Ψ is also surjective. Hence Ψ is a continuous bijection from the compact space $\widetilde{\mathbb{S}}^2_+$ onto the Hausdorff space $\mathcal{F} \subset \mathbb{R}^3$, and therefore a homeomorphism.

Finally, $RB([0,1]) = \Psi\left(\{q \in \widetilde{\mathbb{S}}^2_+ : q_3 = 0\}\right)$. Since $\{q \in \widetilde{\mathbb{S}}^2_+ : q_3 = 0\} \subsetneq \widetilde{\mathbb{S}}^2_+$, injectivity implies $RB([0,1]) \subsetneq \Psi(\widetilde{\mathbb{S}}^2_+) = \mathcal{F}$. Therefore, $\mathcal{F} \neq RB([0,1])$. *Q.E.D.*

We next introduce notation to define Case 2. For $\alpha \in [0,1]$, recall $RB(\alpha)$ from (56) and $RB([0,1]) \equiv \{RB(\alpha) : \alpha \in [0,1]\}$. Recall the mapping $\pi : \mathbb{R}^3 \to \mathbb{R}^2$, $\pi(e_1, e_2, e_3) = (e_1, e_2)$, and that $\mathcal{E}^A \equiv \pi(\mathcal{E})$. For $c = (c_1, c_2) \in \mathcal{E}^A$, define

$$I(c) \equiv \{t \in \mathbb{R} : (c_1, c_2, t) \in \mathcal{E}\}, \tag{59}$$

so that $d(c)$ in (5) satisfies $d(c) = \min\{|t| : t \in I(c)\}$. Recall the FA-frontier relative to accuracy loss in (6), $\mathcal{F}^A$, and define the Pareto frontier relative to accuracy loss,

$$\mathcal{PF}^A \equiv \left\{c \in \mathcal{E}^A : \nexists\, c' \in \mathcal{E}^A \text{ s.t. } c'_1 \leq c_1,\ c'_2 \leq c_2, \text{ with at least one strict inequality}\right\}. \tag{60}$$

Let $f(c)$ be the unique element of $I(c)$ with minimal absolute value, and define

$$\Pi(c) \equiv (c_1, c_2, f(c)). \tag{61}$$

CASE 2: The feasible set $\mathcal{E}$ crosses the plane $\{\varepsilon_3 = 0\}$, i.e., both $\mathcal{E} \cap \{\varepsilon_3 > 0\} \neq \varnothing$ and $\mathcal{E} \cap \{\varepsilon_3 < 0\} \neq \varnothing$, and either $RB([0,1]) \subset \{\varepsilon_3 > 0\}$ or $RB([0,1]) \subset \{\varepsilon_3 < 0\}$.

THEOREM B.2—FA-frontier in Case 2: Let Assumptions 1 and 2 hold, and $\mathcal{E}$ be in Case 2. Then $\mathcal{PF}^A \subsetneq \mathcal{F}^A$ and $RB([0,1]) \subsetneq \mathcal{F}$.

PROOF: By replacing ε_3 with $-\varepsilon_3$ if necessary, we may assume $RB([0,1]) \subset \{\varepsilon_3 > 0\}$. Define:

$$Z \equiv \pi(\mathcal{E} \cap \{\varepsilon_3 = 0\}),$$
$$\mathcal{F}^{A0} \equiv \left\{c \in Z : \nexists\, c' \in Z \text{ such that } c'_1 \leq c_1,\ c'_2 \leq c_2, \text{ with at least one strict inequality}\right\}.$$

Since $\mathcal{E}$ crosses $\{\varepsilon_3 = 0\}$, convexity implies $\mathcal{E} \cap \{\varepsilon_3 = 0\} \neq \varnothing$, hence $Z \neq \varnothing$. The set Z is compact because $\mathcal{E} \cap \{\varepsilon_3 = 0\}$ is compact and π is continuous.

We first show that $\mathcal{F}^{A0} \neq \varnothing$. Choose $c^0 \in Z$ minimizing the continuous map $c \mapsto c_1 + c_2$ on the compact set Z. If there existed $c' \in Z$ such that $c'_1 \leq c^0_1$ and $c'_2 \leq c^0_2$, with at least one strict inequality, then $c'_1 + c'_2 < c^0_1 + c^0_2$, contradicting the choice of c^0. Thus $c^0 \in \mathcal{F}^{A0}$, so $\mathcal{F}^{A0} \neq \varnothing$.

Next we claim that $\mathcal{F}^{A0} \subseteq \mathcal{F}^A$. Let $c \in \mathcal{F}^{A0}$. Since $c \in Z$, we have $0 \in I(c)$, hence $d(c) = 0$. Suppose, towards a contradiction, that $c \notin \mathcal{F}^A$. Then there exists $c' \in \mathcal{E}^A$ such that $c'_1 \leq c_1$, $c'_2 \leq c_2$, $d(c') \leq d(c) = 0$, with at least one strict inequality. Because $d \geq 0$ on $\mathcal{E}^A$, it follows that $d(c') = 0$. Hence $0 \in I(c')$, so $c' \in Z$. Moreover, the strict inequality cannot occur only in the d-coordinate, because $d(c') = d(c) = 0$. Therefore at least one of $c'_1 < c_1$ or $c'_2 < c_2$ holds, contradicting $c \in \mathcal{F}^{A0}$. This proves $\mathcal{F}^{A0} \subseteq \mathcal{F}^A$.

We now show that $\mathcal{PF}^A \cap Z = \varnothing$. Let $c \in \mathcal{PF}^A$. By Lemma B.4, $\Pi(c) \in \Pi(\mathcal{PF}^A) = RB([0,1])$. Because $\mathcal{E}$ is in Case 2, $\Pi(c)_3 > 0$. In particular, $\Pi(c)_3 \neq 0$, so $0 \notin I(c)$, and hence $c \notin Z$. Therefore $\mathcal{PF}^A \cap Z = \varnothing$.

Since $\mathcal{F}^{A0} \subseteq Z$, we obtain $\mathcal{F}^{A0} \cap \mathcal{PF}^A = \varnothing$. Together with $\mathcal{F}^{A0} \subseteq \mathcal{F}^A$ and $\mathcal{F}^{A0} \neq \varnothing$, this yields $\mathcal{F}^A \setminus \mathcal{PF}^A \neq \varnothing$.

On the other hand, $\mathcal{PF}^A \subseteq \mathcal{F}^A$. Indeed, if $c \notin \mathcal{F}^A$, then there exists $c' \in \mathcal{E}^A$ such that $c'_1 \leq c_1$, $c'_2 \leq c_2$, $d(c') \leq d(c)$, with at least one strict inequality. If $c'_1 = c_1$ and $c'_2 = c_2$, then $c' = c$, which forces $d(c') = d(c)$, contradicting the requirement that at least one inequality be strict. Hence at least one of $c'_1 < c_1$ or $c'_2 < c_2$ holds, so $c \notin \mathcal{PF}^A$. Thus $\mathcal{PF}^A \subseteq \mathcal{F}^A$.

Combining the last two results, we conclude that

$$\mathcal{PF}^A \subsetneq \mathcal{F}^A.$$

Finally, since $\pi \circ \Pi = \mathrm{id}_{\mathcal{E}^A}$, the map Π is injective. Therefore

$$\Pi(\mathcal{PF}^A) \subsetneq \Pi(\mathcal{F}^A).$$

Using equivalence (i) $\Leftrightarrow$ (iv) in Theorem 2.2 and Lemma B.4, we obtain

$$RB([0,1]) = \Pi(\mathcal{PF}^A) \subsetneq \Pi(\mathcal{F}^A) = \mathcal{F}.$$

This proves $RB([0,1]) \subsetneq \mathcal{F}$. *Q.E.D.*

LEMMA B.3: The projection $\mathcal{E}^A$ is compact, strictly convex, and has a nonempty interior. For each $\alpha \in [0,1]$, let $c(\alpha) \equiv \arg\min_{c \in \mathcal{E}^A} \left[\alpha c_1 + (1-\alpha)c_2\right]$. Then the Pareto frontier in (60) satisfies $\mathcal{PF}^A = \{c(\alpha) : \alpha \in [0,1]\}$.

PROOF: Since π is continuous and $\mathcal{E}$ is compact, $\mathcal{E}^A$ is compact. To see that $\mathcal{E}^A$ has a nonempty interior, choose $x \in \mathrm{int}(\mathcal{E})$. There exists $r > 0$ such that $\mathbb{B}(x,r) \subset \mathcal{E}$. Then the open disk in $\mathbb{R}^2$ centered at $\pi(x)$ with radius r is contained in $\pi(\mathbb{B}(x,r)) \subset \mathcal{E}^A$. Hence $\mathrm{int}(\mathcal{E}^A) \neq \varnothing$. To prove strict convexity of $\mathcal{E}^A$, let $c^0, c^1 \in \mathcal{E}^A$ be distinct, and choose lifts $\varepsilon^0, \varepsilon^1 \in \mathcal{E}$ with $\pi(\varepsilon^i) = c^i$. Since $c^0 \neq c^1$, we have $\varepsilon^0 \neq \varepsilon^1$. For any $\lambda \in (0,1)$, strict convexity of $\mathcal{E}$ gives $m = \lambda\varepsilon^0 + (1-\lambda)\varepsilon^1 \in \mathrm{int}(\mathcal{E})$. Choose $r > 0$ such that $\mathbb{B}(m,r) \subset \mathcal{E}$. As above, the open disk in $\mathbb{R}^2$ centered at $\pi(m) = \lambda c^0 + (1-\lambda)c^1$ with radius r lies in $\mathcal{E}^A$. Therefore $\lambda c^0 + (1-\lambda)c^1 \in \mathrm{int}(\mathcal{E}^A)$, so $\mathcal{E}^A$ is strictly convex.

Now fix $\alpha \in [0,1]$ and write $L_\alpha(c) = \alpha c_1 + (1-\alpha)c_2$. By strict convexity of $\mathcal{E}^A$, L_α has a unique minimizer $c(\alpha)$ on $\mathcal{E}^A$. We first show that $c(\alpha) \in \mathcal{PF}^A$. Suppose not. Then there exists $c' \in \mathcal{E}^A$ such that

$$c'_1 \leq c(\alpha)_1, \qquad c'_2 \leq c(\alpha)_2,$$

with at least one strict inequality. If $\alpha \in (0,1)$, then $L_\alpha(c') < L_\alpha(c(\alpha))$, contradicting the minimality of $c(\alpha)$. If $\alpha = 0$, then $c'_2 \leq c(\alpha)_2$; if the inequality is strict, we again contra-

dict minimality, while if $c_2' = c(\alpha)_2$, then $c' \neq c(\alpha)$ and c' is a second minimizer of L_0, contradicting uniqueness. The case $\alpha = 1$ is analogous. Thus $c(\alpha) \in \mathcal{PF}^A$.

Conversely, let $x \in \mathcal{PF}^A$. Set $K \equiv \mathcal{E}^A - x$ and $O \equiv \{u \in \mathbb{R}^2 : u_1 < 0, u_2 < 0\}$. Since $x \in \mathcal{PF}^A$, there is no point of $\mathcal{E}^A$ strictly south-west of x, hence $K \cap O = \varnothing$. The set K is compact and convex, while O is open and convex. By the standard strong separation theorem, there exist $q = (q_1, q_2) \in \mathbb{R}^2 \setminus \{0\}$ and $\gamma \in \mathbb{R}$ such that $q^\intercal u \geq \gamma > q^\intercal v$, $\forall u \in K$, $\forall v \in O$. Since $0 \in K$, we obtain $\gamma \leq 0$. We claim that $q_1, q_2 \geq 0$. If $q_1 < 0$, then for $t > 0$, $v_t = (-t, -1) \in O$, but $q^\intercal v_t = -tq_1 - q_2 \to +\infty$ as $t \to +\infty$, contradicting $q^\intercal v_t < \gamma$. Thus $q_1 \geq 0$. Similarly $q_2 \geq 0$.

Next, for $n \geq 1$, let $v_n = \left(-\frac{1}{n}, -\frac{1}{n}\right) \in O$. Since $q_1, q_2 \geq 0$, we have $q^\intercal v_n \to 0^-$. Because $\gamma > q^\intercal v_n$ for every n, it follows that $\gamma \geq 0$. Hence $\gamma = 0$. Therefore, $q^\intercal(y - x) \geq 0 \; \forall y \in \mathcal{E}^A$, or equivalently $q^\intercal y \geq q^\intercal x \; \forall y \in \mathcal{E}^A$. Since $q \neq 0$ and $q_1, q_2 \geq 0$, we have $q_1 + q_2 > 0$. Set $\alpha := \frac{q_1}{q_1 + q_2} \in [0, 1]$. Then q is a positive multiple of $(\alpha, 1 - \alpha)$, so x minimizes L_α on $\mathcal{E}^A$. By uniqueness, $x = c(\alpha)$.

This proves $\mathcal{PF}^A = \{c(\alpha) : \alpha \in [0, 1]\}$. *Q.E.D.*

LEMMA B.4: Let $c(\alpha) \equiv \arg\min_{c \in \mathcal{E}^A}[\alpha c_1 + (1 - \alpha)c_2]$. For every $\alpha \in [0, 1]$, $RB(\alpha) = \Pi(c(\alpha))$. Consequently, $RB([0, 1]) = \Pi(\mathcal{PF}^A) = \mathcal{PF}$, with $\mathcal{PF}$ in (13).

PROOF: Fix $\alpha \in [0, 1]$, and write $L_\alpha(x) := \alpha x_1 + (1 - \alpha)x_2$. By strict convexity of $\mathcal{E}$, the minimizer $RB(\alpha)$ of L_α on $\mathcal{E}$ is unique. By Lemma B.3, the minimizer $c(\alpha)$ of L_α on $\mathcal{E}^A$ is also unique. Since L_α depends only on the first two coordinates, $\pi(RB(\alpha))$ minimizes L_α on $\mathcal{E}^A$. By uniqueness, $\pi(RB(\alpha)) = c(\alpha)$.

We claim that the fiber $I(c(\alpha))$ (with $I(\cdot)$ in (59)) is a singleton. If not, pick $t_1 \neq t_2$ in $I(c(\alpha))$. Then both $(c(\alpha), t_1) \in \mathcal{E}$ and $(c(\alpha), t_2) \in \mathcal{E}$ have the same L_α-value, namely the minimum of L_α on $\mathcal{E}$. This contradicts the uniqueness of $RB(\alpha)$. Hence, $I(c(\alpha)) = \{RB(\alpha)_3\}$, so $f(c(\alpha)) = RB(\alpha)_3$, and therefore $RB(\alpha) = \Pi(c(\alpha))$.

Taking the union over $\alpha \in [0, 1]$ and using Lemma B.3, we obtain

$$RB([0, 1]) = \{RB(\alpha) : \alpha \in [0, 1]\} = \{\Pi(c(\alpha)) : \alpha \in [0, 1]\} = \Pi(\mathcal{PF}^A).$$

To obtain the last equivalence, we first show $\varepsilon \in \mathcal{PF} \iff \pi(\varepsilon) \in \mathcal{PF}^A$. Suppose first that $\varepsilon \in \mathcal{PF}$, and set $c := \pi(\varepsilon)$. If $c \notin \mathcal{PF}^A$, then there exists $c' \in \mathcal{E}^A$ such that $c'_1 \leq c_1$, $c'_2 \leq c_2$, with at least one strict inequality. Choose $\varepsilon' \in \mathcal{E}$ with $\pi(\varepsilon') = c'$. Then $\varepsilon'_1 \leq \varepsilon_1$, $\varepsilon'_2 \leq \varepsilon_2$, with at least one strict inequality, contradicting $\varepsilon \in \mathcal{PF}$. Therefore $\pi(\varepsilon) \in \mathcal{PF}^A$.

Conversely, suppose $\pi(\varepsilon) = c \in \mathcal{PF}^A$. If $\varepsilon \notin \mathcal{PF}$, then there exists $\varepsilon' \in \mathcal{E}$ such that $\varepsilon'_1 \leq \varepsilon_1$, $\varepsilon'_2 \leq \varepsilon_2$, with at least one strict inequality. Setting $c' := \pi(\varepsilon')$, we obtain $c'_1 \leq c_1$, $c'_2 \leq c_2$, with at least one strict inequality, contradicting $c \in \mathcal{PF}^A$. Hence $e \in \mathcal{PF}$. Thus, $e \in \mathcal{PF} \Leftrightarrow \pi(e) \in \mathcal{PF}^A$. It follows that $\mathcal{PF} = \Pi(\mathcal{PF}^A)$. *Q.E.D.*

CASE 3: The feasible set $\mathcal{E}$ crosses the plane $\{\varepsilon_3 = 0\}$, i.e., both $\mathcal{E} \cap \{\varepsilon_3 > 0\} \neq \varnothing$ and $\mathcal{E} \cap \{\varepsilon_3 < 0\} \neq \varnothing$, and $RB([0,1]) \cap \{\varepsilon_3 = 0\} \neq \varnothing$.

REMARK B.1: Under Assumptions 1-2, when $\mathcal{E}$ is in Case 3 we can have $\mathcal{PF} = \mathcal{F}$. In research in progress, we investigate whether $\mathcal{PF} = \mathcal{F}$ always in Case 3.

B.2. *Lemmas And Auxiliary Results for Sections 3-4*

Recall that $\mathcal{F}_k$ denotes the σ-algebra generated by all the data except the ones with indices in $\mathcal{I}_k$.

LEMMA B.5: Let Assumption 4 and 5 hold. Then, there exist $\widehat{\delta}_k : \mathcal{X}_2 \to \mathbb{R}^+$ and an event $\widehat{E}_k \in \mathcal{F}_k$ such that (i) $\widehat{\delta}_k$ conditional on $\mathcal{F}_k$ is nonstochastic, (ii) $\mathbb{P}(\widehat{E}_k^c) = o(1)$, (iii) conditional on $\widehat{E}_k$ we have that $\widehat{\boldsymbol{\mu}}_k \in \mathbb{B}_\epsilon(\mu_r, \mu_b)$, $||\Delta\widehat{\boldsymbol{\theta}}_k(X_i) - \Delta\boldsymbol{\theta}(X_i)|| \leq \widehat{\delta}_k(X_{2,i})$ and $||\widehat{\boldsymbol{\pi}}_k(X_i) - \boldsymbol{\pi}(X_i)|| \leq \widehat{\delta}_k(X_{2,i})$, and (iv) $\mathbb{E}[\widehat{\delta}_k(X_{2,i})^2|\mathcal{F}_k] = o_p(n^{-1/2})$.

PROOF: Let $C_F = \sup_{\nu \in \mathbb{R}^{d_\nu}} \sup_{\gamma \in \mathbb{R}^{d_\gamma}} \sup_{\mu \in \mathbb{B}_\epsilon(\mu_r, \mu_b)} ||DF(\nu, \gamma, \mu)||$. Consider the event $\widehat{E}_k = \{||\widehat{\boldsymbol{\mu}}_k - \boldsymbol{\mu}|| < \epsilon\}$, which holds with probability approaching one by definition of $\hat{\boldsymbol{\mu}}_k$. By Assumptions 4 and 5, the mean value inequality, and conditional on $\widehat{E}_k$, we have

$$\left|\left|\left(\widehat{\boldsymbol{\pi}}(X_i) - \boldsymbol{\pi}(X_i)\ ,\ \Delta\widehat{\boldsymbol{\theta}}(X_i) - \Delta\boldsymbol{\theta}(X_i)\right)\right|\right|$$

is lower or equal than

$$C_F\left(||\widehat{\nu}_k(X_{1,i}) - \nu(X_{1,i})|| + ||\widehat{\gamma}_k(X_{2,i}) - \gamma(X_{2,i})|| + ||\widehat{\boldsymbol{\mu}}_k - \mu||\right)\ .$$

By Assumption 5, the previous expression is lower or equal than

$$\widehat{\delta}_k(X_{2,i}) \equiv C_F \left(\sup_{x_1 \in \mathcal{X}_1} ||\widehat{\nu}_k(x_1) - \nu(x_1)|| + ||\widehat{\gamma}_k(X_{2,i}) - \gamma(X_{2,i})|| + ||\widehat{\boldsymbol{\mu}}_k - \mu|| \right) .$$

By Assumption 5 and Jensen's inequality, we have $\mathbb{E}[\widehat{\delta}_k(X_{2,i})^2|\mathcal{F}_k] = o_p(n^{-1/2})$. *Q.E.D.*

LEMMA B.6: Under the assumptions of Theorem 3.2, $A_k = o_p(n^{-1/2})$ uniformly on $q \in \mathbb{S}^2$, with A_k defined in Eq. (47).

PROOF: It is sufficient to prove that $\mathbb{E}[A_k|\mathcal{F}_k, (X_i)_{i\in\mathcal{I}_k}] = o_p(n^{-1/2})$ to conclude that A_k is $o_p(n^{-1/2})$, by Chernozhukov et al. (2018, Lemma 6.1). Define

$$\widetilde{A}_k \equiv \tfrac{1}{n_k} \sum_{i\in\mathcal{I}_k} |q^\intercal \Delta\boldsymbol{\theta}(X_i)| \, \mathbb{1}\{|q^\intercal \Delta\boldsymbol{\theta}(X_i)| \leq |q^\intercal \Delta\widehat{\boldsymbol{\theta}}_k(X_i) - q^\intercal \Delta\boldsymbol{\theta}(X_i)|\} . \tag{62}$$

Note that $|\mathbb{E}[A_k|\mathcal{F}_k, (X_i)_{i\in\mathcal{I}_k}]| \leq \widetilde{A}_k$; therefore, it is sufficient to show that $\widetilde{A}_k$ is $o_p(n^{-1/2})$ uniformly on $q \in \mathbb{S}^2$.

By Cauchy-Schwartz and Lemma B.5, we have $|q^\intercal \Delta\widehat{\boldsymbol{\theta}}(X_i) - q^\intercal \Delta\boldsymbol{\theta}(X_i)| \leq \widehat{\delta}_k(X_{2,i})$ conditional on $\widehat{E}_n$, where $\widehat{\delta}_k$ and $\widehat{E}_n$ are as in Lemma B.5. Since $\mathbb{P}(\widehat{E}_k^c) = o(1)$, it is sufficient to show that $\mathbb{E}[\widetilde{A}_k \mathbb{1}\{\widehat{E}_K\}|\mathcal{F}_k] = o_p(n^{-1/2})$ to conclude $\widetilde{A}_k$ is $o_p(n^{-1/2})$. We verify this next:

$$\begin{aligned}
\mathbb{E}[\widetilde{A}_k \mathbb{1}\{\widehat{E}_K\}|\mathcal{F}_k] &= \mathbb{E}\left[|q^\intercal \Delta\boldsymbol{\theta}(X_i)| \, \mathbb{1}\{|q^\intercal \Delta\boldsymbol{\theta}(X_i)| \leq |q^\intercal \Delta\widehat{\boldsymbol{\theta}}_k(X_i) - q^\intercal \Delta\boldsymbol{\theta}(X_i)|\} \mathbb{1}\{\widehat{E}_K\}|\mathcal{F}_k \right] \\
&\overset{(1)}{\leq} \mathbb{E}\left[|q^\intercal \Delta\boldsymbol{\theta}(X_i)| \, \mathbb{1}\{|q^\intercal \Delta\boldsymbol{\theta}(X_i)| \leq \widehat{\delta}_k(X_{2,i})\}|\mathcal{F}_k \right] \mathbb{1}\{\widehat{E}_K\} \\
&\overset{(2)}{\leq} \mathbb{E}\left[\widehat{\delta}_k(X_{2,i}) \mathbb{E}\left[\mathbb{1}\{|q^\intercal \Delta\boldsymbol{\theta}(X_i)| \leq \widehat{\delta}_k(X_{2,i})\}|\mathcal{F}_k, X_{2,i} \right] |\mathcal{F}_k \right] \mathbb{1}\{\widehat{E}_K\} \\
&\overset{(3)}{\leq} C_\delta \mathbb{E}\left[\widehat{\delta}_k(X_{2,i})^2 |\mathcal{F}_k \right] \\
&\overset{(4)}{=} o_p(n^{-1/2})
\end{aligned}$$

where (1) holds by Lemma B.5 and since $\mathbb{1}\{\widehat{E}_K\} \in \mathcal{F}_k$, (2) holds since $\widehat{\delta}_k(X_{2,i})$ is non-stochastic conditional on $\mathcal{F}_k$ and $X_{2,i}$, (3) holds by Assumption 4 and since $\mathbb{1}\{\widehat{E}_K\} \leq 1$,

and (4) holds by Lemma B.5. Finally, note that $\mathbb{P}(\widehat{E}_k^c) = o(1)$ and (4) hold uniformly on $q \in \mathbb{S}^2$ since these objects do not depend on q. *Q.E.D.*

LEMMA B.7: Under the assumptions of Theorem 3.2, $B_k = o_p(n^{-1/2})$ uniformly on $q \in \mathbb{S}^2$, with B_k defined in Eq. (48).

PROOF: We can write $B_k = B_{k,1} + B_{k,2}$, where

$$B_{k,1} = \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} q^\intercal \Delta \mathbf{L}_i Z_i \left(\frac{1}{\widehat{\boldsymbol{\pi}}_k(X_i)} - \frac{1}{\boldsymbol{\pi}(X_i)} \right) \left(\mathbb{1}\{q^\intercal \Delta \widehat{\boldsymbol{\theta}}(X_i) > 0\} - \mathbb{1}\{q^\intercal \Delta \boldsymbol{\theta}(X_i) > 0\} \right)$$

$$B_{k,2} = \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \left(\frac{q^\intercal \Delta \widehat{\mathbf{L}}_i Z_i - q^\intercal \Delta \mathbf{L}_i Z_i}{\widehat{\boldsymbol{\pi}}(X_i)} \right) \left(\mathbb{1}\{q^\intercal \Delta \widehat{\boldsymbol{\theta}}(X_i) > 0\} - \mathbb{1}\{q^\intercal \Delta \boldsymbol{\theta}(X_i) > 0\} \right)$$

As in the proof of Lemma B.6, it is sufficient to show that $\mathbb{E}[B_{k,1}|\mathcal{F}_k, (X_i)_{i \in \mathcal{I}_k}]$ and $B_{k,2}$ are both $o_p(n^{-1/2})$ uniformly on $q \in \mathbb{S}^2$. Define

$$\widetilde{B}_{k,1} = \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} |\widehat{\boldsymbol{\pi}}_k(X_i) - \boldsymbol{\pi}(X_i)| \mathbb{1}\{|q^\intercal \Delta \boldsymbol{\theta}(X_i)| \leq |q^\intercal \Delta \widehat{\boldsymbol{\theta}}_k(X_i) - q^\intercal \Delta \boldsymbol{\theta}(X_i)|\}$$

Note that $|\mathbb{E}[B_{k,1}|\mathcal{F}_k, (X_i)_{i \in \mathcal{I}_k}]| \leq \widetilde{B}_{k,1}$; therefore, it is sufficient to show that $\widetilde{B}_{k,1}$ is $o_p(n^{-1/2})$ uniformly on $q \in \mathbb{S}^2$ to conclude the same for $\mathbb{E}[B_{k,1}|\mathcal{F}_k, (X_i)_{i \in \mathcal{I}_k}]$. By similar arguments used in the proof of Lemma B.6 and by relying on Lemma B.5, we conclude

$$\mathbb{E}[\widetilde{B}_{k,1} \mathbb{1}\{\widehat{E}_k\} | \mathcal{F}_k] \leq C_\delta \mathbb{E}\left[\widehat{\delta}_k(X_{2,i})^2 | \mathcal{F}_k \right] = o_p(n^{-1/2}) \,,$$

where $\widehat{E}_k$ and $\widehat{\delta}_k$ are as in Lemma B.5.

Recall the definitions $\mathcal{M}^r = \operatorname{diag}\left(\left[\frac{1}{\mu_r}, 0, \frac{1}{\mu_r} \right] \right)$ and $\mathcal{M}^b = \operatorname{diag}\left(\left[0, \frac{1}{\mu_b}, -\frac{1}{\mu_b} \right] \right)$, with sample analogs denoted by $\widehat{\mathcal{M}}^r$ and $\widehat{\mathcal{M}}^b$, respectively, where $\widehat{\mu}_g = n^{-1} \sum_{i=1}^n \mathbb{1}\{G_i = g\}$. By the next identity

$$q^\intercal \Delta \widehat{\mathbf{L}}_i - q^\intercal \Delta \mathbf{L}_i = \sum_{g \in \{r,b\}} (\widehat{\mathcal{M}}^g q - \mathcal{M}^g q)^\intercal \Delta \mathbf{L}_i \mu_g \mathbb{1}\{G_i = g\} \,,$$

we can write

$$B_{k,2} = \sum_{g \in \{r,b\}} (\widehat{\mathcal{M}}^g q - \mathcal{M}^g q)^\intercal B_{k,2}^g \,,$$

where

$$B^g_{k,2} = \tfrac{1}{n_k} \sum_{i \in \mathcal{I}_k} \left(\tfrac{\Delta \mathbf{L}_i \mu_g \mathbb{1}\{G_i = g\} Z_i}{\widehat{\boldsymbol{\pi}}_k(X_i)} \right) \left(\mathbb{1}\{q^\intercal \Delta \widehat{\boldsymbol{\theta}}_k(X_i) > 0\} - \mathbb{1}\{q^\intercal \Delta \boldsymbol{\theta}(X_i) > 0\} \right) , \quad g \in \{r, b\}$$

We conclude that $\widehat{\mathcal{M}}^g q - \mathcal{M}^g q = O_p(n^{-1/2})$ uniformly on $q \in \mathbb{S}^2$ by the Central Limit Theorem and Assumption 1. Therefore, it is sufficient to show that $B^g_{k,2} = o_p(1)$ uniformly on $q \in \mathbb{S}^2$ to conclude that $B_{k,2} = o_p(n^{-1/2})$ uniformly on $q \in \mathbb{S}^2$. Define

$$\widetilde{B}^g_{k,2} = \tfrac{1}{n_k} \sum_{i \in \mathcal{I}_k} C \mathbb{1}\{|q^\intercal \Delta \boldsymbol{\theta}(X_i)| \leq |q^\intercal \Delta \widehat{\boldsymbol{\theta}}_k(X_i) - q^\intercal \Delta \boldsymbol{\theta}(X_i)|\}$$

where C is a constant based only on c_1, c_2, and ϵ (constants from Assumptions 1 and 4) such that $|\mathbb{E}[B^g_{k,2} | \mathcal{F}_k]| \leq \widetilde{B}^g_{k,2}$. By similar arguments used in the proof of Lemma B.6 and by relying on Lemma B.5, we conclude

$$\mathbb{E}[\widetilde{B}^g_{k,2} \mathbb{1}\{\widehat{E}_k\} | \mathcal{F}_k] \leq C_\delta \mathbb{E}\left[\widehat{\delta}_k(X_{2,i}) | \mathcal{F}_k \right] = o_p(n^{-1/4}) ,$$

which is sufficient to complete the proof of the lemma. *Q.E.D.*

LEMMA B.8: Let Assumption 4 and 5 hold. Then,

$$n_k^{-1/2} \sum_{i \in \mathcal{I}_k} |\widehat{\boldsymbol{\alpha}}_k(q; X_i) - \boldsymbol{\alpha}(q; X_i)|^2 = o_p(1)$$

uniformly on $q \in \mathbb{S}^2$.

PROOF: We can write

$$n_k^{-1/2} \sum_{i \in \mathcal{I}_k} |\widehat{\boldsymbol{\alpha}}_k(q; X_i) - \boldsymbol{\alpha}(q; X_i)|^2 \leq \widetilde{A}_k + B_k + C_k$$

where

$$\widetilde{A}_k = n_k^{-1/2} \sum_{i \in \mathcal{I}_k} |q^\intercal \Delta \boldsymbol{\theta}(X)| \mathbb{1}\{|q^\intercal \Delta \boldsymbol{\theta}(X_i)| \leq |q^\intercal \Delta \widehat{\boldsymbol{\theta}}_k(X_i) - q^\intercal \Delta \boldsymbol{\theta}(X_i)|\}$$

$$B = n_k^{-1/2} \sum_{i \in \mathcal{I}_k} |q^\intercal \widehat{\boldsymbol{\theta}}_{0,k}(X) - q^\intercal \boldsymbol{\theta}_0(X_i)|^2$$

$$C = n_k^{-1/2} \sum_{i \in \mathcal{I}_k} |q^\intercal \Delta\widehat{\boldsymbol{\theta}}_k(X) - q^\intercal \Delta\boldsymbol{\theta}(X)|$$

In the proof of Lemma B.6, we conclude $\widetilde{A}_k = o_p(1)$ uniformly on $q \in \mathbb{S}^2$, with $\widetilde{A}_k$ defined in Eq. (62). Finally, note that B_k and C_k both are $o_p(1)$ uniformly on $q \in \mathbb{S}^2$ by Cauchy-Schwartz and Assumption 5. *Q.E.D.*

LEMMA B.9: Under the assumptions of Theorem 3.2, then

$$\sup_{q \in \mathbb{S}^2} n^{1/2} \left| \widetilde{h}_{\mathcal{E}}(q, \widetilde{\mathbf{L}}, \widehat{\boldsymbol{\eta}}) - \tfrac{1}{n} \sum_{i=1}^{n} \left(\tfrac{W_i}{\overline{W}} \right) \zeta_i^*(q; \mathbf{L}, \boldsymbol{\eta}) \right| = o_p(1) \tag{63}$$

and

$$\sup_{q \in \mathbb{S}^2} n^{1/2} \left| \widetilde{\varepsilon}^* - \tfrac{1}{n} \sum_{i=1}^{n} \left(\tfrac{W_i}{\overline{W}} \right) \xi_i^*(a; \mathbf{L}, \boldsymbol{\eta}) \right| = o_p(1) \tag{64}$$

PROOF: We use the approach presented in the proof of Theorem 3.2. For concreteness, we present the strategy for the proof of Eq. (63), since it is analog for Eq. (64).

Let us denote by $\widetilde{\mathcal{M}}_i = \operatorname{diag}\left(\left[1 - \frac{\mathbb{1}\{G_i=r\}}{\widetilde{\mu}_r}, 1 - \frac{\mathbb{1}\{G_i=b\}}{\widetilde{\mu}_b}, \frac{\mathbb{1}\{G_i=b\}}{\widetilde{\mu}_b} - \frac{\mathbb{1}\{G_i=r\}}{\widetilde{\mu}_r} \right] \right)$ the sample analog of $\mathcal{M}_i$, where μ_g is replaced by $\widetilde{\mu}_g = n^{-1} \sum_{i=1}^{n} \left(\frac{W_i}{\overline{W}} \right) \mathbb{1}\{G_i = g\}$. Recall the definitions $\mathcal{M}^r = \operatorname{diag}\left(\left[\frac{1}{\mu_r}, 0, \frac{1}{\mu_r} \right] \right)$ and $\mathcal{M}^b = \operatorname{diag}\left(\left[0, \frac{1}{\mu_b}, -\frac{1}{\mu_b} \right] \right)$, with the following sample analogs denoted by $\widetilde{\mathcal{M}}^r$ and $\widetilde{\mathcal{M}}^b$, respectively, where we replaced μ_g by $\widetilde{\mu}_g$. Note that $\widetilde{\mathcal{M}}^r q - \mathcal{M}^r q = O_p(n^{-1/2})$ uniformly on $q \in \mathbb{S}^2$ by the Central Limit Theorem, Assumption 1, and the definition of the independent random variables $\{W_i : 1 \leq i \leq n\}$.

We adapt the arguments used in Part 1 of proof of Theorem 3.1 to prove (63). There, we divided the proof in three claims. Their analogs statements appears next:

claim 1: $\sup_{q \in \mathbb{S}^2} n^{1/2} \left| \widetilde{h}_{\mathcal{E}}(q, \widetilde{\mathbf{L}}, (\Delta\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}_0)) - \widetilde{h}_{\mathcal{E}}(q, \widetilde{\mathbf{L}}, (\Delta\boldsymbol{\theta}, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}_0)) \right| = o_p(1)$

To show claim 1, we can adapt Lemma B.6 by using that $\{W_i : 1 \leq i \leq n\}$ is independent of the data. Similarly, we can adapt Lemma B.7 by using also that $\widetilde{\mathcal{M}}^r q - \mathcal{M}^r q = O_p(n^{-1/2})$ uniformly on $q \in \mathbb{S}^2$.

claim 2: $\sup_{q \in \mathbb{S}^2} n^{1/2} \left| \widetilde{h}_{\mathcal{E}}(q, \widetilde{\mathbf{L}}, (\Delta\boldsymbol{\theta}, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}_0)) - \frac{1}{n} \sum_{i=1}^{n} \left(\frac{W_i}{\overline{W}} \right) \zeta_i^*(q; \mathbf{L}, (\Delta\boldsymbol{\theta}, \widehat{\boldsymbol{\pi}}, \widehat{\boldsymbol{\theta}}_0)) \right| = o_p(1)$

To show claim 2, note that $n^{-1}\sum_{i=1}^{n}\left(\frac{W_i}{\overline{W}}\right)\widetilde{\mathcal{M}}_i = 0$. We use this to write

$$\widetilde{h}_{\mathcal{E}}(q,\widetilde{\mathbf{L}},\widehat{\boldsymbol{\eta}}) = \tfrac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i}{\overline{W}}\right)\left[\zeta_i(q;\widetilde{\mathbf{L}},(\boldsymbol{\theta},\widehat{\boldsymbol{\pi}},\widehat{\boldsymbol{\alpha}}^q)) + (\widetilde{\mathcal{M}}_i q)^{\intercal}\nabla_q h_{\mathcal{E}}(q)\right].$$

We then use that $\{W_i : 1 \leq i \leq n\}$ is independent of the data and $\widetilde{\mathcal{M}}^r q - \mathcal{M}^r q = O_p(n^{-1/2})$ uniformly on $q \in \mathbb{S}^2$ to conclude.

claim 3: $\sup_{q\in\mathbb{S}^2} n^{1/2}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\frac{W_i}{\overline{W}}\right)\left(\zeta_i^*(q;\mathbf{L},(\Delta\boldsymbol{\theta},\widehat{\boldsymbol{\pi}},\widehat{\boldsymbol{\theta}}_0)) - \zeta_i^*(q;\mathbf{L},\boldsymbol{\eta})\right)\right| = o_p(1)$

Finally, to show claim 3, we use that $\{W_i : 1 \leq i \leq n\}$ is independent of the data. *Q.E.D.*

LEMMA B.10: Let $\widehat{\varepsilon}^*$ be as in Eq. (31). Under the assumptions of Theorem 3.2, then

$$\sqrt{n}\,(\widehat{\varepsilon}^* - \varepsilon^*) = \mathbb{G}[\xi_i^*(a^*;\mathbf{L},\boldsymbol{\eta})]\,,$$

where $\xi_i^*(a^*;\mathbf{L},\boldsymbol{\eta}) = \xi_i(a^*;\mathbf{L},\boldsymbol{\eta}) + \sum_{g\in\{r,b\}}\boldsymbol{\alpha}^{e,g}\,(1 - \mathbb{1}\{G_i = g\}/\mu_g)$, $\xi(a^*;\mathbf{L},\boldsymbol{\eta})$ is as in Eq. (29), and $\boldsymbol{\alpha}^{e,g} = \mathbb{E}\big[\mathbb{1}\{G_i = g\}\,(\mathbf{L}_{0,i} + a(X_i)(\mathbf{L}_{1,i} - \mathbf{L}_{0,i}))\,\frac{Z_i}{\boldsymbol{\pi}(X_i)}\big]$.

PROOF: We first prove that $\sqrt{n}\left(\widehat{\varepsilon}^* - n^{-1}\sum_{i=1}^{n}\xi_i^*(a^*;\mathbf{L},\boldsymbol{\eta})\right) = o_p(1)$. The proof follows by similar arguments as in part 1 of the proof of Theorem 3.2; therefore, it is omitted. We then conclude by the Central Limit Theorem. *Q.E.D.*