Chiseling: Powerful and Valid Subgroup Selection via Interactive Machine Learning

Nathan Cheng[1], Asher Spector[2], and Lucas Janson[1]

[1]Department of Statistics, Harvard University
[2]Department of Statistics, Stanford University

Abstract

In regression and causal inference, *controlled subgroup selection* aims to identify, with inferential guarantees, a subgroup (defined as a subset of the covariate space) on which the average response or treatment effect is above a given threshold. E.g., in a clinical trial, it may be of interest to find a subgroup with a positive average treatment effect. However, existing methods either lack inferential guarantees, heavily restrict the search for the subgroup, or sacrifice efficiency by naive data splitting. We propose a novel framework called *chiseling* that allows the analyst to *interactively* refine and test a candidate subgroup by iteratively shrinking it. The sole restriction is that the shrinkage direction only depends on the points outside the current subgroup, but otherwise the analyst may leverage any prior information or machine learning algorithm. Despite this flexibility, chiseling controls the probability that the discovered subgroup is null (e.g., has a non-positive average treatment effect) under minimal assumptions: for example, in randomized experiments, this inferential validity guarantee holds under only bounded moment conditions. When applied to a variety of simulated datasets and a real survey experiment, chiseling identifies substantially better subgroups than existing methods with inferential guarantees.

1 Introduction

1.1 Motivation

When a drug is tested in a clinical trial, the primary question of interest is typically: "Will people benefit on average?" This leads to inference on the drug's population average treatment effect. But a strictly more informative question is: "*Who* will benefit from the drug?" It could be that a drug does not benefit the population on average but provides substantial average benefit to a subgroup. The task of finding such a subgroup is called *subgroup selection* and has a long history of study, but methods for subgroup selection face the fundamental challenge of using the same data for both identifying a subgroup and performing statistical inference on that subgroup. Thus, the analyst either faces a difficult tradeoff between discovery and validation, or else imposes structural assumptions (such as a parametric

model) to address the efficiency loss. However, such approaches typically rule out the use of complex but powerful modern machine learning algorithms whose theoretical properties are poorly understood.

The statistical problem of subgroup selection—which can also be framed as the discovery and evaluation of a treatment policy—has causal applications beyond clinical trials: Under what conditions does a crop benefit from a fertilizer? Who will be swayed by a political ad? Who will benefit from a proposed economic policy? Further applications arise in non-causal regression: Who is likely to get cancer? Who is likely to purchase a new product? What proteins are likely to have high binding affinity to a target molecule? This paper's proposed method will be equally equipped to answer such questions.

We emphasize the critical importance of not just identifying a subgroup but also providing statistical assurance of that subgroup's merit. There are myriad methods that can be used to *propose* a subgroup, but especially when benefiting subgroups are small or nonexistent in the population, these methods can be unreliable and prone to identifying invalid subgroups (i.e. subgroups where the average benefit is zero or negative). Confidence in the validity of the subgroup is crucial when substantial cost (e.g. in terms of time, money, or reputation) is associated with the identification of invalid subgroups.

1.2 Our contribution

In this paper, we describe a new methodological framework for *controlled subgroup selection* in causal inference and regression. Controlled subgroup selection, formally defined in Section 2, is the task of identifying a data-dependent subgroup (a region of covariate space) endowed with a high probability guarantee that it is valid (i.e. that its average response is positive). In causal inference, the identification of such a subgroup immediately yields a *treatment policy* (the policy that treats all units in the subgroup) with positive benefit over the policy that treats no one. Our key insight is the ability to improve a subgroup by shrinking it exclusively based on the data outside that subgroup, leaving *all* the data inside the subgroup untouched and usable for inference. By deploying this insight incrementally and interactively, we allow the analyst to leverage any machine learning algorithm or domain knowledge to craft the best subgroup they can while ensuring high power to test the efficacy of the learned subgroup. We call this approach *chiseling* and explain the analogy at the end of Section 3.2.

Our approach does not limit the complexity of the discovered subgroup, and the analyst may leverage state-of-the-art, black-box machine learning (ML) methods to flexibly search the covariate space without fear of violating the statistical guarantee. Our method produces valid subgroups under very weak assumptions. In the case of regression with binary outcomes, our method is finite-sample valid and exact under only the assumption that the data points are i.i.d. For causal inference, in randomized experiments, we guarantee asymptotic validity under only mild, dimension-free moment assumptions (in addition to standard causal identification assumptions). We show in extensive simulations and in a re-analysis of a real randomized experiment that our method produces consistently, and sometimes dramatically, better subgroups than existing methods with comparable validity guarantees.

1.3 Related work

Because subgroup selection is relevant across many disciplines, the literature is extensive. We focus in the main text on giving a bird's-eye view of works that provide similar frequentist inferential guarantees to ours *without* naive data splitting (as data splitting is general and can be applied to any subgroup selection procedure), that are comparably flexible, and are valid under comparably weak assumptions. We defer to Appendix A discussion of works that focus on the design of selection criteria without in-sample evaluation (e.g. Manski, 2004; Kitagawa and Tetenov, 2018; Lipkovich et al., 2011; Athey and Imbens, 2016; Spiess et al., 2023), that rely on a parametric model for validity (e.g. Ma and Huang, 2017; Wan et al., 2024), that utilize a Bayesian approach (e.g. Dixon and Simon, 1991; Berger et al., 2014; Schnell et al., 2016; Hill, 2011), and that leverage ideas in conformal inference (e.g. Medarametla and Candès, 2021; Jin and Candes, 2023). We also defer to Appendix A discussion of works in sequential hypothesis testing (e.g. Wald, 1945; Lan and DeMets, 1983; Lai et al., 2014), with which we share some high level similarities, but review works in interactive hypothesis testing, which is more closely related to our method, in this subsection.

Methods that select and evaluate subgroups from a structured class. Some works take as given a prespecified, structured class of subgroups such as a finite set or low-complexity policy class (e.g. with finite Vapnik–Chervonenkis dimension). Examples of the latter include parametric classes, stochastic mixtures of a finite set of policies, and the set of subgroups that arise from continuously thresholding an *a priori* fixed prognostic marker (which may be a machine learning score trained out of sample). Inference is then accomplished by either simultaneous inferences over all candidate subgroups (Bonetti, 2004; Song and Pepe, 2004; Imai and Li, 2023; Chernozhukov et al., 2025) or inference on a data-dependent estimand such as the one corresponding to the largest estimated effect (Guo and He, 2020; Andrews et al., 2023). Central to these approaches is the *a priori* fixed class of subgroups (usually with some low dimensional structure), thus limiting the use of machine learning methods unless those models are trained out of sample (i.e. via data splitting).

CATE-based approaches. There is a line of work on efficient estimation of the CATE (Künzel et al., 2019; Nie and Wager, 2020; Kennedy, 2023). Some works produce confidence bands for the conditional average treatment effect (CATE) function, either via efficient estimation (Wager and Athey, 2018; Ritzwoller and Syrgkanis, 2024; Armstrong and Shen, 2015) or leveraging other structural assumptions such as smoothness or isotonicity (Reeve et al., 2023; Müller et al., 2024). CATE inference may be implausible absent these assumptions, which become particularly formidable in high dimensions. A compromise is to consider inference on a *calibrated projection* of the CATE, either onto a small subset of variables (Fan et al., 2020; Zimmert and Lechner, 2019; Semenova and Chernozhukov, 2020) or a parametric family (Cai et al., 2010; Cai et al., 2011). However, the complexity of the projection class can alter the finite-sample behavior of these approaches, and thus in these works there is intrinsically a tradeoff between flexibility and validity.

Repeated data splitting. Data splitting (Cox, 1975) is a flexible and essentially assumption-free solution to controlled subgroup selection. However, Fithian et al. (2017) observe that naively applying data splitting is inadmissible in many situations because it discards more information than is used by selection. Likewise, we empirically improve over naive data splitting under similarly weak assumptions. Chernozhukov et al. (2018) leverage

repeated data splitting to reduce dependence on a single split, but this yields inference for an *aggregated* estimand that does not correspond to the effect for any one subgroup. Cross-validation, though we have not seen it applied to subgroup selection, can be used to validate a selected subgroup under stability assumptions on the selection algorithm (Bayle et al., 2020; Austern and Zhou, 2020). Jia et al. (2024) split the data into folds to *sequentially* learn and validate an improving sequence of policies, but they require stability conditions similar to those of cross-validation. Stability can be difficult to guarantee for complex algorithms, high-dimensional data, and data with weak signals. For demonstrations of the inferential invalidity of cross-validation in simple settings, see Bates et al. (2023) and Bayle et al. (2025). Because our method does not require stability assumptions, it is more broadly valid.

Interactive hypothesis testing. Interactive hypothesis testing was introduced in Lei and Fithian (2018) and extended in Lei et al. (2020) and Chao and Fithian (2021) for interactive control of the false discovery rate. It has been extended to tests of the global null (Duan et al., 2020b), family-wise error rate control (Duan et al., 2020a), and tests of distributional equality (Duan et al., 2022). At a high level, these interactive procedures all rely on a construction that iteratively reveals some information to the analyst while keeping the remaining information hidden. Notably, Duan et al. (2024), working in a causal setting, describe an interactive procedure that produces a subset of units of a given data set while controlling the expected proportion of units for whom the treatment was not effective. However, the paper concerns inferences on the actual experimented-upon units and thus does not yield a usable treatment policy, while our work concerns guarantees for *distributional properties* of subregions of the covariate space such as average effects. Broadly, our work draws inspiration from the interactive literature while deviating in the important regard that previous works concern inferences on properties of a finite sample such as the proportion of nulls or proportion of benefiting individuals in the sample. We are motivated by the spirit of interactive procedures but require significantly different technical tools to solve our problem.

2 Problem statement: controlled subgroup selection

Although we are primarily motivated by problems in causal inference, we start with a regression formulation. We observe n i.i.d. data points $(X_1, Y_1), ..., (X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}$. Let $\mathcal{D} := (X_i, Y_i)_{i=1}^n$ and let (X, Y) denote a generic sample from the distribution of (X_i, Y_i) which is independent of $\mathcal{D}$. We seek to identify a region $\mathcal{R} \subseteq \mathcal{X}$ such that $\mathbb{E}[Y \mid X \in \mathcal{R}]$ is strictly larger than some user-specified cutoff $\mu_{\text{cut}} \in \mathbb{R}$.

In the context of causal inference, let $Y'(1), Y'(0) \in \mathbb{R}$ denote a unit's potential outcomes for treatment and control, respectively, and let $X \in \mathcal{X}$ denote that unit's covariates. We seek to find a subset of the covariate space where the average treatment effect for that subgroup is above μ_{cut}. In a randomized controlled trial (RCT) with treatment indicator W, under standard causal assumptions, we can inverse probability weight (IPW) each unit's potential outcomes to produce a regression problem of the above form. Letting

$$Y = \frac{W}{\mathbb{E}[W \mid X]} Y'(1) - \frac{1 - W}{\mathbb{E}[1 - W \mid X]} Y'(0), \tag{1}$$

4

then Y is observable and $\mathbb{E}[Y \mid X]$ equals the conditional average treatment effect (CATE) $\mathbb{E}[Y'(1) - Y'(0) \mid X]$. In particular,

$$\mathbb{E}[Y \mid X \in \mathcal{R}] = \mathbb{E}[Y'(1) - Y'(0) \mid X \in \mathcal{R}].$$

Hence, the regression formulation we will consider throughout this paper in some sense subsumes the causal one. We will occasionally use causal language to build intuition, in which case it should be understood that the regression has been converted from an underlying causal problem via Equation (1). Note that this problem statement and our method can also use the more efficient augmented IPW (AIPW) transformation, but since this does not preserve the i.i.d. structure of the data, it requires additional care; see Appendix D.5 for details.

For a fixed subgroup $\mathcal{R}$, define the *subgroup mean*

$$\mu(\mathcal{R}) := \mathbb{E}[Y \mid X \in \mathcal{R}],$$

let $\mu(\emptyset)$ be undefined, and let the statement "$\mu(\emptyset) \leq \mu_{\text{cut}}$" evaluate as false. This is also the *subgroup average treatment effect* when Y comes from the IPW transformation. By slight abuse of notation, we will let $\mu(x) := \mu(\{x\}) = \mathbb{E}[Y \mid X = x]$ for $x \in \mathcal{X}$. Given a nominal level $\alpha \in [0, 1]$, the precise inferential validity guarantee we seek to provide is that our data-dependent subgroup $\mathcal{R}$ satisfies

$$\mathbb{P}(\mu(\mathcal{R}) \leq \mu_{\text{cut}}) \leq \alpha. \tag{2}$$

To expand on this criteria: when $\mathcal{R}'$ is fixed, a classical subgroup analysis reports $\mathcal{R} = \mathcal{R}'$ if $H_0 : \mu(\mathcal{R}') \leq \mu_{\text{cut}}$ is rejected and $\mathcal{R} = \emptyset$ otherwise. Then Equation (2) exactly reduces to describing a Type I error guarantee for H_0. By analogy, we generically refer to Equation (2) as Type I error control, even when $\mathcal{R}$ is data-dependent. When $\mu_{\text{cut}} = 0$, this guarantees that we erroneously report a subgroup whose average effect is nonpositive no more than α of the time (and where a vacuous discovery $\mathcal{R} = \emptyset$ is not considered an error). We call this task *controlled subgroup selection*, since we seek to select a subgroup from the data while controlling the Type I error.

Subject to the Type I error constraint, we measure the quality of a subgroup selection procedure by how well it maximizes the *expected utility* of the reported subgroup for some chosen notion of utility. Abstractly, if $\mathcal{U}(\mathcal{R})$ measures the utility of $\mathcal{R}$, then we seek to maximize the expected utility,

$$\mathbb{E}[\mathcal{U}(\mathcal{R})], \tag{3}$$

subject to Equation (2). Different notions of utility may be more or less appropriate depending on the inferential goals and the task at hand. One natural choice is

$$\mathcal{U}(\mathcal{R}) := \mathbb{E}[(Y - \mu_{\text{cut}})\mathbb{1}\{X \in \mathcal{R}\}] = (\mu(\mathcal{R}) - \mu_{\text{cut}})\rho(\mathcal{R}) \tag{4}$$

where $\rho(\mathcal{R}) := \mathbb{P}(X \in \mathcal{R})$. This is maximized by the upper level set of the conditional mean function: $\mathcal{R}^* := \{x \in \mathcal{X} : \mathbb{E}[Y \mid X = x] > \mu_{\text{cut}}\}$. In the policy learning literature μ_{cut} is often taken to be 0, in which case treating units in $\mathcal{R}^*$ is an optimal policy (Appendix C.1). By construction, "$\mu(\mathcal{R}^*) \leq \mu_{\text{cut}}$" is false. Using Equation (4) as the utility rewards regions

5

that are more similar to $\mathcal{R}^*$, which is precisely the subgroup of units whose predictable (i.e. X-conditional) means exceed μ_{cut}.

Other natural utility functions include $\mathcal{U}(\mathcal{R}) := \mathbb{1}\{\mathcal{R} \neq \emptyset\}$ and $\mathcal{U}(\mathcal{R}) := \rho(\mathcal{R})$. The expectation of the former is *power*, and the expectation of the latter is *expected probability mass*. However, for the remainder of the paper, we will let $\mathcal{U}(\cdot)$ be defined as in Equation (4) and simply refer to it as "the utility" unless otherwise specified, since this is one of the most common objectives in causal inference and its choice will help illustrate the key ideas in our paper.

Notation and conventions. Let $[n] := \{1, ..., n\}$ and $[n]_0 := \{0, ..., n\}$. Let Ω denote the sample space of the base probability measure. All random variables (including the data $(X_i, Y_i)_{i=1}^n$) and σ-algebras are defined with respect to Ω. We will abbreviate "almost surely" as "a.s." For random variables A and B, we say that $A \leq B$ almost surely conditional on an event E if $\mathbb{P}(\{A \leq B\} \cap E) = \mathbb{P}(E)$, and similarly with "$\leq$" signs replaced with "$=$" signs. We let $\Phi(\cdot)$ denote the standard normal CDF. We will write $(x)_+ = \max\{0, x\}$, sometimes use the wedge notation for minimum, i.e. $a \wedge b = \min\{a, b\}$, and will let the minimum of an empty set be ∞. We assume that all functions are measurable throughout. We will sometimes use language such as "choose a function $f(\cdot)$ based on $\mathcal{F}$" where $\mathcal{F}$ is a σ-algebra. Formally, what we mean by this is that $f(\cdot)$ is an $\mathcal{F}$-measurable random function.

3 Methodology

3.1 Overview of our approach

The challenge of controlled subgroup selection is that the data needs to be used to both (i) identify the candidate subgroup and (ii) perform statistical inference on that subgroup. Data splitting accomplishes this under extremely weak assumptions, and we improve upon data splitting by revealing information in an incremental fashion, allowing us to leverage more information before definitively selecting any particular subgroup. Algorithm 1 summarizes the full approach. Below, we also give a high-level overview of our main ideas.

High-level overview. Chiseling proceeds by initializing the region under consideration as the full population $\mathcal{R} = \mathcal{X}$ and interweaving two basic operations: (1) a step that shrinks $\mathcal{R}$, and (2) a step that chooses a fraction of α to "spend" on testing $H_0 : \mu(\mathcal{R}) \leq \mu_{\text{cut}}$, stopping and reporting $\mathcal{R}$ if a rejection is made.

The key idea behind (1) is that as long as the shrinkage directions are chosen based on data outside of $\mathcal{R}$, the subset of data that lies in $\mathcal{R}$ is *untarnished* (defined in Section 3.2) in the sense that it is an i.i.d. sample of units from $\mathcal{R}$. This makes it simple to test properties of $\mathcal{R}$ even though $\mathcal{R}$ is a complex, data-dependent object. The key idea behind (2) is that it is possible to characterize the dependence between the tests that we design for $H_0 : \mu(\mathcal{R}) \leq \mu_{\text{cut}}$ as $\mathcal{R}$ evolves throughout the procedure. This lets us non-conservatively test multiple regions until a rejection is made via *conditionally valid tests* (introduced in Section 3.5).

A feature of our approach is its immense flexibility: Type I error is controlled regardless of how the analyst shrinks the region (any machine learning method can be used) or allocates α, as long as decisions are made only based on information allowed by the protocol. Throughout,

---

Algorithm 1 Chiseling

  Input: dataset $\mathcal{D} = (X_i, Y_i)_{i=1}^n$, cutoff μ_{cut}, and nominal Type I error level α

 1: Initialize region $\mathcal{R}_0 \leftarrow \mathcal{X}$ and revealed information $\mathcal{F}_0 \leftarrow \{\emptyset, \Omega\}$
 2: **for** $t = 0, ..., m$ **do**
 3:    Choose level $\alpha_t \in [0, 1]$ based on $\mathcal{F}_t$, ensuring valid α-budget   ▷ Definition 3.5
 4:    Perform conditionally valid level-α_t test ϕ_t of $\mathcal{H}_t : \mu(\mathcal{R}_t) \leq \mu_{\text{cut}}$ ▷ Definition 3.6
 5:    **if** ϕ_t rejects **then**
 6:     **Return:** $\mathcal{R}_t$
 7:    **end if**
 8:    Obtain next region $\mathcal{R}_{t+1}$ by shrinking $\mathcal{R}_t$ based on $\mathcal{F}_t$ ▷ Algorithm 2
 9:    See revealed data points: $\mathcal{F}_{t+1} \leftarrow \sigma(\mathcal{F}_t, (X_i, Y_i)_{i:X_i \in \mathcal{R}_t \setminus \mathcal{R}_{t+1}})$
10: **end for**
11: **If Line 6 never executes, return:** $\emptyset$

---

we will suggest guidelines to navigate this flexibility, including ways to construct interpretable regions (Section 4.5).

3.2 Untarnished subgroups

When a subgroup $\mathcal{R}$ is fixed *a priori*, inferences about $\mathcal{R}$ are straightforward because $(X_i, Y_i)_{i:X_i \in \mathcal{R}}$ behaves like an i.i.d. sample from the distribution of $(X, Y) \mid X \in \mathcal{R}$. Thus, any standard inferential procedure can be applied to the subgroup sample $(X_i, Y_i)_{i:X_i \in \mathcal{R}}$ to yield inferences about $\mathcal{R}$. For instance, the sample mean of $(Y_i)_{i:X_i \in \mathcal{R}}$ is unbiased for $\mathbb{E}[Y \mid X \in \mathcal{R}]$, and the usual confidence intervals derived from the central limit theorem (CLT) are asymptotically valid. Our primary insight is that the key property that the distribution of points inside $\mathcal{R}$ is not "distorted" can also hold for a *data-dependent region*, conditional on having selected the region, if we construct it carefully. We begin with a useful definition.

Definition 3.1 (Untarnished subgroup). *Let $\mathcal{R} \subseteq \mathcal{X}$ be a (possibly data-dependent) subgroup, and let $\mathcal{F}$ be a σ-algebra. We say that $\mathcal{R}$ is* untarnished *with respect to $\mathcal{F}$ if $\mathcal{R}$ and $n(\mathcal{R}) := |\{i : X_i \in \mathcal{R}\}|$ are $\mathcal{F}$-measurable, and the distribution of the subsample $(X_i, Y_i)_{i:X_i \in \mathcal{R}}$ conditional on $\mathcal{F}$ is the same as that of an independent sample of $n(\mathcal{R})$ data points drawn i.i.d. from the distribution of $(X, Y) \mid X \in \mathcal{R}$.*

Formalizing what we just discussed, an *a priori* fixed $\mathcal{R}$ is untarnished with respect to $\mathcal{F} := \sigma(n(\mathcal{R}))$. Regions selected by data splitting can also be thought of as untarnished (see Section 3.4). To see how else untarnishedness arises, consider the following thought experiment. Let $\mathcal{R}$ be fixed *a priori*. Since the data points outside of $\mathcal{R}$ are conditionally independent of the data points inside of it, $\mathcal{R}$ is untarnished with respect to $\mathcal{F} := \sigma((X_i, Y_i)_{i:X_i \notin \mathcal{R}})$. Thus, even upon scrutinizing $(X_i, Y_i)_{i:X_i \notin \mathcal{R}}$, an analyst may still treat $(X_i, Y_i)_{i:X_i \in \mathcal{R}}$ like an i.i.d. sample from $(X, Y) \mid X \in \mathcal{R}$. For instance, in Figure 1, the blue points constitute our state of knowledge after revealing the points outside the non-data-dependent region $\mathcal{R} = (0, \infty) \subseteq \mathbb{R}$. Studying them, we may be dissatisfied with $\mathcal{R}$. After all, if the optimal subgroup $\{x \in \mathbb{R} : \mathbb{E}[Y \mid X = x] > 0\}$ is our target, a simple linear regression (the solid blue
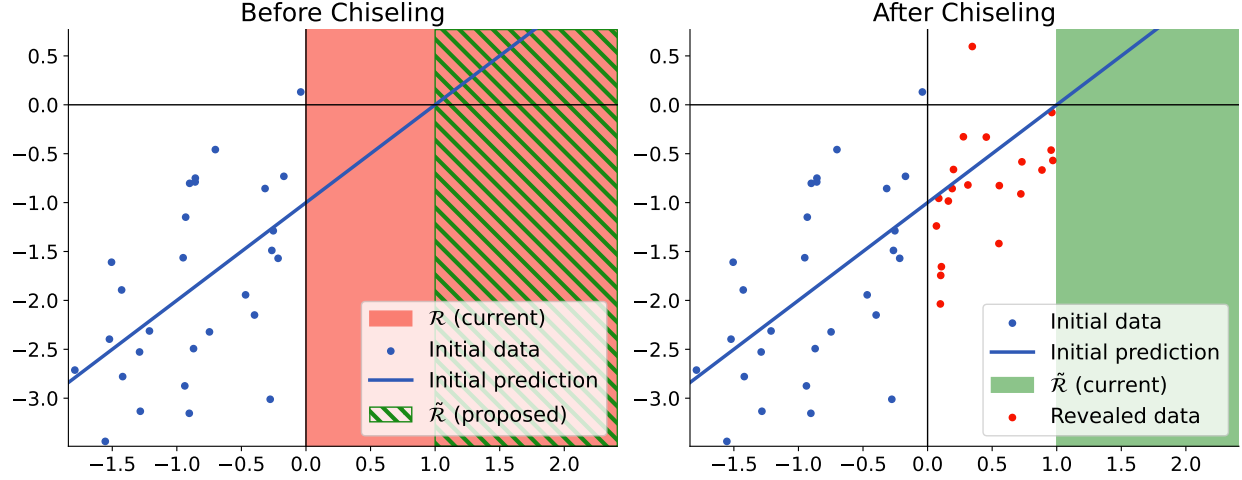
7

Figure 1: A thought experiment using a toy example: the blue line represents the least squares fit through the initially revealed blue points. After shrinking the red region $\mathcal{R} = (0, \infty)$ to the green region $\tilde{\mathcal{R}} = (1, \infty)$, the red points are revealed.

line) suggests that $\tilde{\mathcal{R}} := (1, \infty)$ is better than $\mathcal{R}$. Our key insight is this: we may switch the focus of our inference to $\tilde{\mathcal{R}}$, and $\tilde{\mathcal{R}}$ remains untarnished even though it was determined in a data-dependent way, because it was only determined using points outside of $\mathcal{R}$. Thus, the further subsetted sample $(X_i, Y_i)_{i:X_i \in \tilde{\mathcal{R}}}$ may be used as if it were an i.i.d. sample from $\tilde{\mathcal{R}}$.

In this example, untarnishedness does *not* hinge on the correctness of the linear extrapolation. But we will be rewarded with a better region and generally more powerful inferences if our modeling assumptions are accurate. Furthermore, $\tilde{\mathcal{R}}$ is untarnished with respect to $\tilde{\mathcal{F}} := \sigma(\mathcal{F}, (X_i, Y_i)_{i:X_i \in \mathcal{R} \setminus \tilde{\mathcal{R}}})$ because the newly excluded points (the red points) are conditionally independent of the points inside of $\tilde{\mathcal{R}}$. Thus, if we remain dissatisfied after inspecting $(X_i, Y_i)_{i:X_i \notin \tilde{\mathcal{R}}}$, for instance finding that our new least squares fit has shifted substantially, we may repeat the above process to further shrink $\tilde{\mathcal{R}}$. But since we cannot reverse this process, it seems prudent to shrink as little as possible at each step in order to avoid overshrinking. Indeed, our construction in the next subsection shrinks infinitesimally until new data is revealed.

We call our approach *chiseling*, in analogy with how a sculptor reduces their starting material to produce a final artwork, refining their artistic choices as they interact with the medium, and moreover, *committing* to any reductions they make, since chiseled material cannot be restored to the main body. Later, we will explore why the chiseling framework is beneficial, including a simple numerical demonstration in Section 3.4. But first, the next subsection describes in more detail how chiseling is operationalized and some of its properties.

3.3 Chiseling subroutine: region shrinking

The level sets of functions allow us to conveniently parameterize region shrinking. Let $\mathcal{R}$ be the current region and untarnished with respect to $\mathcal{F}$. For any function $f : \mathcal{X} \to \mathbb{R}$, let $\tilde{\mathcal{R}}(z) := \mathcal{R} \cap \{x \in \mathcal{X} : f(x) > z\}$. It helps to imagine $\tilde{\mathcal{R}}(z)$ shrinking from $\mathcal{R}$ to $\emptyset$ as z "sweeps" continuously from $-\infty$ to ∞. Where should z stop? There is no loss of generality

in letting z sweep no further than $\min_{i:X_i\in\mathcal{R}} f(X_i)$, the level required to reveal at least one new data point, since to shrink more we may simply repeat this procedure. But also, we may wish to cap z at some value c even if no data points fall in $\mathcal{R} \setminus \tilde{\mathcal{R}}(c)$ (e.g. to ensure we only reveal points with negative predicted treatment effects; see Section 3.4). Thus, we define the next region as

$$\tilde{\mathcal{R}} := \mathcal{R} \cap \left\{ x \in \mathcal{X} : f(x) > c \wedge \left(\min_{i:X_i\in\mathcal{R}} f(X_i) \right) \right\}. \tag{5}$$

We allow $f(\cdot)$ and c to be chosen in an arbitrary fashion by the analyst as long as they are chosen based on $\mathcal{F}$ (i.e. they are $\mathcal{F}$-measurable). We also allow the choice $c = \infty$, which removes the role of c from the above construction. Essentially, this shrinks $\mathcal{R}$ along the contours defined by $f(\cdot)$ until either new data is revealed or c is reached. For maximizing utility in causal inference, a natural idea is to let $f(\cdot)$ be an estimate of the CATE using the information in $\mathcal{F}$, as then units with the lowest predicted CATEs are revealed first, in principle leaving more of the optimal region intact. We will soon see that this choice is empirically effective (Section 3.4), though further constraints may be imposed if the analyst desires more interpretable regions (Section 4.5).

Algorithm 2 Chiseling subroutine (region shrinking)

    Input: dataset $\mathcal{D} = (X_i, Y_i)_{i=1}^n$, current region $\mathcal{R}$, current information $\mathcal{F}$

1: Analyst chooses $f : \mathcal{X} \to \mathbb{R}$ and $c \in \mathbb{R} \cup \{\infty\}$ based on $\mathcal{F}$
2: **Return:** $\tilde{\mathcal{R}} := \mathcal{R} \cap \{x \in \mathcal{X} : f(x) > c \wedge (\min_{i:X_i\in\mathcal{R}} f(X_i))\}$

Algorithm 2 summarizes the construction we have just described, which is a subroutine of Algorithm 1. We call $f(\cdot)$ the *scoring function* and c the *cap*, and we say that Algorithm 2 *chisels* $\mathcal{R}$ to produce a new region $\tilde{\mathcal{R}} \subseteq \mathcal{R}$.[1] The following theorem formalizes the property described in the thought experiment visualized by Figure 1.

Theorem 3.2 (Chiseling preserves untarnishedness). *Suppose that $\mathcal{R}$ is untarnished with respect to $\mathcal{F}$ (Definition 3.1). Let $\tilde{\mathcal{R}}$ be the output of Algorithm 2 applied to $(\mathcal{D}, \mathcal{R}, \mathcal{F})$, and define $\tilde{\mathcal{F}} := \sigma(\mathcal{F}, (X_i, Y_i)_{i:X_i\in\mathcal{R}\setminus\tilde{\mathcal{R}}})$. Then $\tilde{\mathcal{R}}$ is untarnished with respect to $\tilde{\mathcal{F}}$.*

A proof is given in Appendix D.1.2. We now describe how to chain Algorithm 2 to produce a nested sequence of regions $(\mathcal{R}_t)_{t=0}^{\infty}$ while accumulating information the analyst can use for selection. Section 3.5 will describe how to select one of these regions while controlling Type I error. We assume without loss of generality that $\mathcal{R}_0 = \mathcal{X}$, the entire covariate space, and let $\mathcal{F}_0 := \{\emptyset, \Omega\}$, the trivial σ-algebra. Thus, $\mathcal{R}_1 \subseteq \mathcal{R}_0$ must be obtained by chiseling $\mathcal{R}_0$ using only prior information (see Section 4.1 for a discussion of this point). Then recursively define for $t = 1, 2, ...,$

$$\begin{aligned} \mathcal{R}_t &:= \text{ the output of Algorithm 2 applied to } (\mathcal{D}, \mathcal{R}_{t-1}, \mathcal{F}_{t-1}), \\ \mathcal{F}_t &:= \sigma(\mathcal{F}_{t-1}, (X_i, Y_i)_{i:X_i\in\mathcal{R}_{t-1}\setminus\mathcal{R}_t}). \end{aligned} \tag{6}$$

---

[1]While we call our entire shrinking and testing procedure (Algorithm 1) chiseling, we will also refer to the act of shrinking a region (via Algorithm 2) as chiseling.

Observing that $\mathcal{R}_0$ is trivially untarnished with respect to $\mathcal{F}_0$ since it is not data-dependent, the first part of the following corollary follows directly from recursively applying Theorem 3.2. The second part is proved in Appendix D.1.3.

Corollary 3.3 (Sequentially chiseled regions)**.** *For any fixed t, $\mathcal{R}_t$ is untarnished with respect to $\mathcal{F}_t$. If ν is an a.s. finite stopping time with respect to $(\mathcal{F}_t)_{t=0}^{\infty}$, then $\mathcal{R}_\nu$ is untarnished with respect to $\mathcal{F}_\nu$, the corresponding stopped σ-algebra.*

That is, the subsample $(X_i, Y_i)_{i:X_i \in \mathcal{R}_t}$ behaves like an i.i.d. sample of size $n(\mathcal{R}_t)$ from $(X, Y) \mid X \in \mathcal{R}_t$, conditionally on $\mathcal{F}_t$, even if t is a stopping time ν. While Corollary 3.3 applies to an infinite sequence of regions, practically it is only necessary to apply chiseling a finite number of times to produce a finite sequence of regions. Thus, unless stated otherwise we will fix an m and restrict our attention to the first $m+1$ steps and the sequence $(\mathcal{R}_t, \mathcal{F}_t)_{t=0}^{m}$.

Remark 3.4 (Auxiliary randomness)**.** *Corollary 3.3 still holds if the analyst leverages auxiliary randomness (randomness that is independent of the data) when making decisions such as selecting $f(\cdot)$ and c. For instance, the analyst may leverage machine learning methods with inherent randomness such as random forests or stochastic gradient descent. In this case, we simply condition on the auxiliary randomness and omit it from our discussion.*

In summary, as long as we only shrink the current region, we may alter the direction of this shrinkage at any time, in any way, and particularly in light of new information conferred by data points that are excluded from the region. Corollary 3.3 ensures that $\mathcal{R}_t$ is untarnished with respect to $\mathcal{F}_t$, and thus the samples within it are suitable for use in standard inferential procedures. For any stopping time ν, Corollary 3.3 paves the way for inferences on $\mathcal{R}_\nu$. In fact, something much more general is possible: in Section 3.5, we will show that it is possible to tightly "split" one's α error budget over many regions in the sequence $(\mathcal{R}_t)_{t=0}^{m}$, so that instead of testing $H_0 : \mu(\mathcal{R}_\nu) \leq 0$, regions may be tested successively until a rejection is made. Still, even the simplest approach of using samples in $(X_i, Y_i)_{i:X_i \in \mathcal{R}_\nu}$ for inference on $\mathcal{R}_\nu$ yields immediate improvements over data splitting, as we highlight in the next subsection.

3.4 Chiseling is more efficient than data splitting

To understand why chiseling improves upon data splitting, we first describe data splitting in a way that reveals a direct comparison. Let $p \in (0, 1)$ and let $\mathcal{M}$ denote a random subset of $[n]$ of size $\lceil (1-p)n \rceil$ so that there are $\lfloor pn \rfloor$ data points in $\mathcal{D}_{\text{train}} := (X_i, Y_i)_{i \notin \mathcal{M}}$. A standard use of data splitting would build an estimate $\hat{\mu}(\cdot)$ of $\mu(\cdot)$ via machine learning on $\mathcal{D}_{\text{train}}$, define $\mathcal{R}_{\text{ds}} := \{x \in \mathcal{X} : \hat{\mu}(x) > \mu_{\text{cut}}\}$, and test $H_0 : \mu(\mathcal{R}_{\text{ds}}) \leq \mu_{\text{cut}}$ by running a one-sided t-test using the independent test subsample $(X_i, Y_i)_{i \in \mathcal{M}:X_i \in \mathcal{R}_{\text{ds}}}$. But another way to view data splitting is by unraveling the construction of $\mathcal{R}_{\text{ds}}$ into a shrinking sequence of upper level sets of $\hat{\mu}(\cdot)$ that terminates at $\mathcal{R}_{\text{ds}}$. In particular, if we consider the process $\mathcal{X} = \mathcal{R}_0 \supseteq \mathcal{R}_1 \supseteq ... \supseteq \mathcal{R}_{\text{ds}}$ where

$$\mathcal{R}_t = \left\{ x \in \mathcal{X} : \hat{\mu}(x) > \mu_{\text{cut}} \wedge \left(\min_{i \in \mathcal{M}:X_i \in \mathcal{R}_{t-1}} \hat{\mu}(X_i) \right) \right\}$$

then we see that $\mathcal{R}_{\text{ds}}$ is simply a *special case* of chiseling where in Algorithm 2 we use $f(\cdot) = \hat{\mu}(\cdot)$—fit only using $\mathcal{D}_{\text{train}}$—and $c = \mu_{\text{cut}}$ at every shrinking step, and where $(X_i, Y_i)_{i \in \mathcal{M}}$
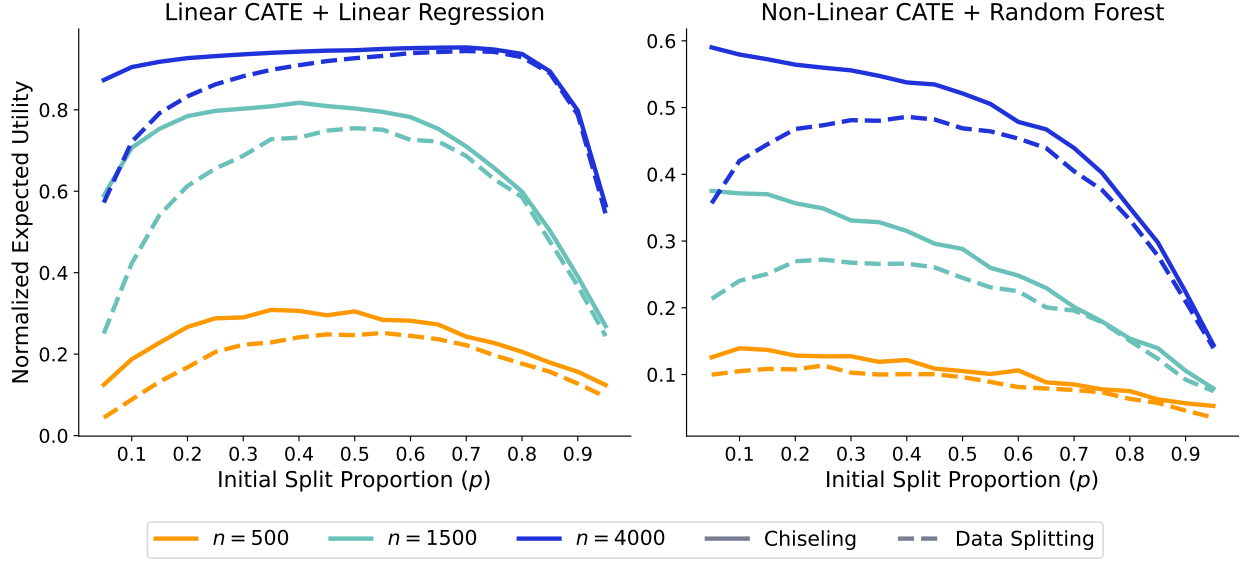
Figure 2: Normalized expected utility of region reported by chiseling versus data splitting, controlling for the same starting information (via initial split proportion) and machine learning method in two simulated RCTs with three different sample sizes n.

is treated as the dataset. Formally, chiseling can also be seen to generalize the full data splitting procedure (including fitting $\hat{\mu}(\cdot)$ to $\mathcal{D}_{\text{train}}$) via an argument that adds an independent, auxiliary dimension to the covariates and chisels in that dimension (see Appendix C.4).

From here, it is easy to see how data splitting can be improved. Corollary 3.3 tells us that at the tth stage, the data that is excluded from $\mathcal{R}_t$ can be used without invalidating the t-test that is run on the final region.[2] Thus, chiseling can follow the exact same steps as above with just one change: at each stage, refit $\hat{\mu}(\cdot)$ to *all* the data that is available, which includes $\mathcal{D}_{\text{train}}$ and the points revealed thus far by shrinking. This implies that, given the same $\mathcal{D}_{\text{train}}$, chiseling will tend to find better regions than data splitting since it iteratively improves its estimate of $\mu(\cdot)$ using more data. As a consequence, chiseling can find better regions with smaller split proportions p, leaving larger sample sizes available for testing.

Figure 2 compares the expected utility of this simple variant of chiseling (see Appendix C.6 for a precise description) to data splitting across a range of initial split proportions p in two simulated RCTs: one where the CATE follows a linear model and the other where the CATE is nonlinear (see Appendix B.2 for details). As expected, for every n and p, chiseling dominates data splitting, and the difference across split proportions p can be quite substantial, e.g., for $n = 1500$ in the second panel, the maximum of chiseling's expected utility curve is 38% higher than that of the data splitting curve. Furthermore, we see that there are typically many choices for p where, for most or all $p' \in (0, 1)$, chiseling with a p initial split outperforms data splitting with a p' initial split.

While leveraging an initial split allows us to directly compare to data splitting, chiseling can also be applied without an initial split if there is useful prior information—we discuss this possibility in Section 4.1. But first, we situate chiseling in an even more general sequential

---

[2] Later, once we have defined conditionally valid tests, one will see that running a t-test on the final region is a particular instantiation of Algorithm 1.

11

testing framework that gives the analyst more flexibility in how regions are tested.

3.5 General framework: sequential testing of chiseled regions

As there is uncertainty in our estimate of the optimal region, we may prefer not to over-commit to testing just a single region, but rather spread power over multiple candidate regions. Though multiple testing can be addressed via the Bonferroni correction, this will be conservative because the data points in $\mathcal{R}_t$ greatly overlap with those in $\mathcal{R}_{t+1}$, creating strong positive dependence between successive tests.[3] Instead, we introduce a framework based on constructing a sequence of hypothesis tests ϕ_t which are *conditionally valid* for null hypotheses $\mathcal{H}_t$ at level α_t. In this paper, $\mathcal{H}_t$ will be the hypothesis that $\mathcal{R}_t$, the tth region produced by chiseling, satisfies $\mu(\mathcal{R}_t) \leq \mu_{\text{cut}}$.

We first present a very general version of this idea. As we will allow the error allocated to each adaptively determined null hypothesis to be chosen in an online manner, we need the following definition.

Definition 3.5 (α-budget)**.** *For a fixed α, we call $(\alpha_t)_{t=0}^m$ a valid α-budget with respect to a filtration $(\mathcal{G}_t)_{t=0}^m$ if α_t is $\mathcal{G}_t$-measurable for all $t = 0, ..., m$ and $1 - \prod_{t=0}^m (1 - \alpha_t) \leq \alpha$ almost surely. If additionally, $1 - \prod_{t=0}^m (1 - \alpha_t) = \alpha$ almost surely, then we call it an exact α-budget.*

When choosing α_t in an online manner, Definition 3.5 can always be satisfied by choosing $\alpha_t \leq 1 - (1 - \alpha) \prod_{s=0}^{t-1} (1 - \alpha_s)^{-1}$ based on $\mathcal{G}_t$.

Next, let us formalize a null hypothesis $\mathcal{H}$ as both a proposition about the data distribution and a random variable that evaluates to 0 if that proposition is true (we are under the null) and 1 otherwise (we are under the alternative). For instance, when $\mathcal{R}$ is random the hypothesis $\mathcal{H} : \mu(\mathcal{R}) \leq \mu_{\text{cut}}$ corresponds to the random variable $\mathcal{H} = \mathbb{1}\{\mu(\mathcal{R}) > \mu_{\text{cut}}\}$. We formalize a test ϕ for $\mathcal{H}$ as a binary random variable that evaluates to 1 if $\mathcal{H}$ is rejected; thus, a false rejection occurs if and only if $\mathcal{H} = 0$ and $\phi = 1$.

Definition 3.6 (Conditionally valid testing sequence)**.** $(\mathcal{H}_t, \alpha_t, \phi_t)_{t=0}^m$ *is a* conditionally valid testing sequence *with respect to a filtration* $(\mathcal{G}_t)_{t=0}^m$ *if for all $t = 0, ..., m$ we have that $\mathcal{H}_t$ and $\alpha_t \in [0, 1]$ are $\mathcal{G}_t$-measurable and the following holds:*

$$\mathbb{P}(\phi_t = 1 \mid \sigma(\mathcal{G}_t, (\phi_s)_{s<t})) \leq \alpha_t \text{ a.s. on the event } \{\mathcal{H}_t = 0 \text{ and } \max_{s<t} \phi_s = 0\}. \quad (7)$$

Additionally, we call $(\mathcal{H}_t, \alpha_t, \phi_t)_{t=0}^m$ *a* conditionally exact testing sequence *with respect to* $(\mathcal{G}_t)_{t=0}^m$ *if Equation (7) holds with equality for all $t = 0, ..., m$.*

In words, Definition 3.6 asks that the random variable ϕ_t specifies a valid level-α_t test of the null hypothesis $\mathcal{H}_t$ *conditional* on all the information in $\mathcal{G}_t$ as well as the fact that no previous test rejected. A simple lemma shows that conditionally valid testing sequences control the overall Type I error rate. Though the result is intuitive, its proof contains subtleties and is presented in Appendix D.2.

---

[3]Nonetheless, the Bonferroni correction constitutes a valid and general approach, and has some other interesting uses that we discuss in Appendix C.16.

Lemma 3.7 (Sequential error control). *Let $(\mathcal{H}_t, \alpha_t, \phi_t)_{t=0}^m$ be a conditionally valid testing sequence and $(\alpha_t)_{t=0}^m$ be a valid α-budget with respect to a filtration $(\mathcal{G}_t)_{t=0}^m$. Define the first rejected index $\tau := \min\{t : \phi_t = 1\}$ with $\tau = \infty$ if the set is empty, and let $\mathcal{H}_\infty = 1$ by convention. Then rejecting $\mathcal{H}_\tau$ controls the Type I error rate:*

$$\mathbb{P}(\mathcal{H}_\tau = 0) \leq \alpha. \tag{8}$$

Furthermore, if $(\mathcal{H}_t, \alpha_t, \phi_t)_{t=0}^m$ is conditionally exact, $(\alpha_t)_{t=0}^m$ is an exact α-budget, and $\mathcal{H}_t = 0$ almost surely for all $t = 0, ..., m$ (i.e. we are under the global null), then Equation (8) holds with equality.

The above lemma ensures that $\mathcal{H}_\tau$, the rejected hypothesis, is a true null with probability no greater than α. We now connect this result to our subgroup selection problem. By convention, let $\mathcal{R}_\infty = \emptyset$ (corresponding to no rejection). Now letting $\mathcal{H}_t := \mathbb{1}\{\mu(\mathcal{R}_t) > \mu_{\text{cut}}\}$, Lemma 3.7 implies that an appropriately designed filtration $(\mathcal{G}_t)_{t=0}^m$, sequence of levels $(\alpha_t)_{t=0}^m$, and collection of tests $(\phi_t)_{t=0}^m$ yields the desired error control,

$$\mathbb{P}(\mu(\mathcal{R}_\tau) \leq \mu_{\text{cut}}) \leq \alpha,$$

where we recall that $\tau := \min\{t : \phi_t = 1\}$. That is, by rejecting the first (largest) region in our sequence of chiseled regions using conditionally valid tests, we satisfy Type I error control.[4]

At this point we have specified Algorithm 1 up to the design of the conditionally valid tests $(\phi_t)_{t=0}^m$, which we will do in the following subsection, and strategies for shrinking the regions and designing α-budgets, which we discuss in Section 4.

3.6   Design of the tests ϕ_t and main validity results

Constructing conditionally valid tests may seem quite challenging. For our problem, however, conditional tests can be readily constructed by leveraging the untarnishedness of chiseled regions. Since by Corollary 3.3 the sample mean of points in a chiseled region $\mathcal{R}_t$ would be unbiased for $\mu(\mathcal{R}_t)$ given $\mathcal{F}_t$ if we did not condition on the previous tests $(\phi_s)_{s<t}$, arguably the most natural test simply thresholds the sample mean of points in the region, i.e.

$$\phi_t := \mathbb{1}\{\bar{\mu}_t > C_t\}$$

where[5]

$$\bar{\mu}_t := \frac{1}{n_t} \sum_{i:X_i \in \mathcal{R}_t} Y_i \quad \text{and} \quad n_t := |\{i : X_i \in \mathcal{R}_t\}|.$$

The challenge is to choose C_t so that the tests are part of a conditionally valid testing sequence, with the main difficulty arising from the fact that conditioning on $\{\max_{s<t} \phi_s =$

---

[4]In Appendix C.15, we describe a variant of chiseling that rejects multiple regions while controlling the family-wise error rate.

[5]If ever $\bar{\mu}_t$ or C_t is undefined, we may either let $\phi_t = 0$, which is trivially conditionally valid, or let $\phi_t \sim \text{Bern}(\alpha_t)$ independently, which is trivially conditionally exact.

0} creates bias. Fortunately, it is tractable to characterize this bias. It turns out that $\max_{s<t} \phi_s = 0$ if and only if $\bar{\mu}_t \leq M_t$ where

$$M_t := \min_{s<t} \left\{ \frac{1}{n_t} \left(n_s C_s - \sum_{i:X_i \in \mathcal{R}_s \setminus \mathcal{R}_t} Y_i \right) \right\}. \tag{9}$$

See Appendix C.7 for a detailed derivation.

First, when Y is binary, we can derive a sequence of exact tests that demonstrates our main idea (we return to the case of general Y in a moment).[6] We make use of the fact that prior to truncation, $n_t \bar{\mu}_t \sim \text{Binomial}(n_t, \mu(\mathcal{R}_t))$, which is stochastically dominated by $\text{Binomial}(n_t, \mu_{\text{cut}})$ under the null. This leads to the following definition and theorem (see Appendix C.9 for a more detailed motivation).

Definition 3.8 (Binary Y critical values). *For $t = 0, ..., m$, let*

$$C_t := \frac{Q(1 - \alpha_t; n_t, \mu_{\text{cut}}, n_t M_t)}{n_t}$$

where M_t is defined by Equation (9), and $Q(q; \tilde{n}, \tilde{\mu}, \tilde{M})$ denotes the qth quantile of a Binomial($\tilde{n}, \tilde{\mu}$) truncated to be no greater than $\tilde{M}$ (see Appendix C.8 for a precise definition).

Theorem 3.9 (Validity and tightness for binary Y). *Assume Y is binary and let $\mathcal{R}_\tau$ be the region produced by Algorithm 1 using tests of the form $\phi_t := \mathbb{1}\{\bar{\mu}_t > C_t\}$ where C_t is defined in Definition 3.8. Then*

$$\mathbb{P}(\mu(\mathcal{R}_\tau) \leq \mu_{\text{cut}}) \leq \alpha.$$

If in addition $\mu(X) = \mu_{\text{cut}}$ a.s. (i.e., we are at the boundary of the global null) and $1 - \prod_{t=0}^{m}(1 - \alpha_t) = \alpha$ almost surely, then the Type I error guarantee is tight in the sense that

$$\mathbb{P}(\mu(\mathcal{R}_\tau) \leq \mu_{\text{cut}}) = \alpha.$$

A proof is given in Appendix D.3.1. In it, we use the important fact that stochastic domination for binomials is preserved under right truncations (Lemma F.6), which is not true for general distributions. Of particular note is the fact that Theorem 3.9 places no restrictions or assumptions on the choice of $f(\cdot)$, c, and α_t beyond what is already stated in Algorithms 1 and 2, providing the user enormous latitude in how they make such choices in order to maximize the utility of the returned subgroup $\mathcal{R}_\tau$ without concern of violating Type I error. Furthermore, the only distributional assumptions we have made are that the dataset $(X_i, Y_i)_{i=1}^{n}$ consists of i.i.d. data points and that Y is binary, yet the validity guarantee is non-asymptotic and tight. It is worth noting that here C_t, which is based on the null distribution truncated at M_t, is deterministically no larger than the $1 - \alpha_t$ quantile of Binomial(n_t, μ_{cut}), and potentially much smaller if M_t is small. The latter is what the critical value would be if we based it on the *untruncated* null distribution (the null distribution conditional only on

---

[6]For causal inference with binary outcomes, Equation (1) does not produce binary Y. One may either use the asymptotic variant described later, or use a finite-sample valid extension to binary potential outcomes described in Appendix C.17.

$\mathcal{F}_t$). Thus, in our case conditioning makes it easier for each test to reject compared to not conditioning.

When the support of Y is unrestricted, we rely on the fact that when n_t is large, $\bar{\mu}_t$ is approximately Gaussian with approximately known variance. We also suppose without loss of generality that $\mu_{\mathrm{cut}} = 0$ in the remainder of this subsection, since for any other μ_{cut} we may simply subtract it from Y to obtain an equivalent problem where $\mu_{\mathrm{cut}} = 0$. The following critical values are motivated by a Gaussian approximation.

Definition 3.10 (General Y critical values). *In increasing order of t for $t = 0, ..., m$, let*

$$C_t := \left(\frac{\Phi^{-1}((1 - \alpha_t) \cdot \Phi(\sqrt{n_t} \cdot \hat{V}_t^{-1/2} M_t))}{\sqrt{n_t} \cdot \hat{V}_t^{-1/2}} \right)_+ \qquad where \qquad \hat{V}_t := \frac{1}{n_t} \sum_{i : X_i \in \mathcal{R}_t} (Y_i - \bar{\mu}_t)^2$$

and M_t is defined by Equation (9). *The numerator of C_t is the $(1 - \alpha_t)$th quantile of a standard normal truncated to be no greater than $\sqrt{n_t} \cdot \hat{V}_t^{-1/2} M_t$.*

To precisely state our asymptotic validity guarantee for Algorithm 1 with general outcomes, we must define our asymptotics. Let $(X_i, Y_i)_{i=1}^{\infty}$ be an infinite stream of i.i.d. samples from the distribution of (X, Y), and let $\mathcal{R}_{\tau}^{(n)}$ be the output of Algorithm 1 obtained by treating the first n data points as the entire dataset $\mathcal{D}$. Here we also allow m, the maximum number of chiseling steps, to vary with n. Our asymptotics will be taken as $n \to \infty$. As with all CLT-based inference, we need some very basic moment assumptions about the data.

Assumption 3.11 (Moment conditions). $\mathbb{E}[|Y|^4] < \infty$ *and* $\mathrm{Var}(Y \mid X) > 0$ *a.s.*

Because our tests are based on normal approximations, we need some algorithmic restrictions in order to keep the total approximation error under control. For instance, we need to ensure that the minimum sample size associated to any test with nonzero α_t is growing. We also need to ensure that any nonzero α_t is not too small, since otherwise the inversion from α_t to quantile is too unstable, and we do rely on the accuracy of the quantiles themselves since they figure into the truncation levels.

Constraint 3.12 (Chiseling constraints). *Fix $p_{\mathrm{prop}}, \alpha_{\min} \in (0, 1)$. Algorithm 1 obeys the following constraints for all t:*

1. Minimum sample size: *If $\alpha_t > 0$, then $n_t/n \geq p_{\mathrm{prop}}$.*

2. Tail error restriction: *If $\alpha_t > 0$, then $\alpha_t \geq \alpha_{\min}$.*

Note that the tail error restriction in Constraint 3.12 implies that there is a K_{reg} such that the number of tested regions is bounded, i.e. $|\{t : \alpha_t > 0\}| \leq K_{\mathrm{reg}}$. We emphasize that Constraint 3.12 is not an assumption but a constraint on how Algorithm 1 is instantiated. It is entirely within the analyst's control to satisfy it irrespective of any unknowns about the problem. We suggest practical ways to fulfill Constraint 3.12 and demonstrate that these choices achieve excellent empirical performance in Section 5. We can now state our validity guarantee for Algorithm 1 for general outcomes.

Theorem 3.13 (Validity for general Y). *Suppose (X, Y) satisfies Assumption 3.11 and that Algorithm 1 satisfies Constraint 3.12. When the sample size is n, let $\mathcal{R}_\tau^{(n)}$ be the region produced by Algorithm 1 using tests of the form $\phi_t := \mathbb{1}\{\bar{\mu}_t > C_t\}$ where C_t is defined in Definition 3.10. Then*

$$\limsup_{n \to \infty} \mathbb{P}\big(\mu(\mathcal{R}_\tau^{(n)}) \le 0\big) \le \alpha.$$

A proof is given in Appendix D.4. We note that Theorem 3.13 does not come with an exactness guarantee. This is a byproduct of the clipping at 0 in Definition 3.10, which is necessary due to an asymptotic subtlety requiring us to disallow C_t from diverging to $-\infty$ (see Appendix C.10 for an extended discussion). Nevertheless, we find chiseling to be essentially exact in our simulations (Section 5.1).

Although we state Theorem 3.13 with a pointwise (i.e. fixed (X, Y) distribution) guarantee for simplicity of exposition, in Appendix D.4 we prove validity in a more general triangular array setting that subsumes the pointwise case, which allows the distribution of (X, Y) to also change with n. In particular, we emphasize that the asymptotic validity of Algorithm 1 does not depend on the dimension of X at all (it could even be infinite), nor on any structural assumption about $Y \mid X$ or the consistency of any estimator of $\mathbb{E}[Y \mid X]$, nor on how the analyst shrinks the regions or allocates the α_t as long as they satisfy the mild Constraint 3.12. In fact, we can construct and work with a generalized procedure that relaxes the constraint that the minimum n_t is proportional to n, allowing it to grow to ∞ at any rate and thus permitting the discovery of regions of arbitrarily small probability mass; this general result's statement is quite a bit more technical, so we defer it to Appendix D.4.

For causal problems, the AIPW estimator is generally preferred to the IPW estimator because it is more statistically efficient. Even though the AIPW transformation does not preserve the i.i.d. structure of the resulting sample since it typically utilizes data-dependent nuisance functions, we nonetheless prove that the AIPW estimator can be validly leveraged to perform chiseling in RCTs. As the setup is more verbose, we defer a detailed discussion and proofs to Appendix D.5.

4 Practical considerations

In Section 3.4, we discussed the choice of $f(\cdot)$ and μ_{cut}. We now address some additional practical considerations: the initialization of chiseling, how to set the levels $(\alpha_t)_{t=0}^m$, subgroup selection when non-responders are not harmed, computational efficiency, and the (optional) construction of interpretable regions. We refrain from making any universal prescriptions since chiseling is a flexible framework that an analyst can leverage in creative ways to validly incorporate domain knowledge. However, we hope that our discussion can serve as a helpful guide to practitioners in a breadth of scenarios.

4.1 Initializing chiseling

When domain knowledge or structural assumptions are completely absent, chiseling may be initialized by splitting off a proportion of the sample (as in Section 3.4). Thus, our simulations in Section 5 will leverage such initial data splits to kickstart chiseling. In practice,

however, a researcher with sufficient understanding of their problem can initialize more powerfully and non-randomly. For instance, known directions of effect modification, mechanistic models, and predictive models trained on observational data or data from earlier clinical trial phases may inform the choice of initial shrinking direction. As new data is revealed by shrinking, the analyst is free to incorporate this information into $f(\cdot)$.

When an initial split—let us say of proportion p—is necessary, how large should one take it to be? For the same reason that it is not easy to answer this for data splitting, it is not easy to answer this for chiseling. However, we make two observations. First, chiseling is effective even when p is small because chiseling will continue to learn on data beyond the initial split. Second, any method that relies on an initial split proportion p can be run at multiple $p_1, ..., p_k$ with Type I error rates $\alpha^{(1)} + ... + \alpha^{(k)} = \alpha$, then combined using the Bonferroni correction. In our case, if $p_1 \leq ... \leq p_k$ and $\mathcal{R}^{(1)}, ..., \mathcal{R}^{(k)}$ are the regions reported by the k different analyses, it is natural to report $\mathcal{R}^{(j^*)}$ where $j^* := \max\{j : \mathcal{R}^{(j)} \neq \emptyset\}$, since among nonempty regions $\mathcal{R}^{(j^*)}$ is learned using the most initial training data. We call this *Bonferroni aggregation* and we use it to aggregate results across different p for chiseling, data splitting, and other competing methods in Section 5 (we find $k = 3$, split proportions $\{0.2, 0.5, 0.8\}$, and equal α splitting to be an effective default choice). Surprisingly, this strategy can sometimes outperform any single choice of p run at the full level α, which we discuss in more detail in Appendix C.12.

4.2 Setting the levels α_t

Suppose we run Algorithm 1 by letting at step t the scoring function $f(\cdot)$ be a machine learning estimate of the conditional mean function $\mu(x) = \mathbb{E}[Y \mid X = x]$ based on $\mathcal{F}_t$ and the cap c be set equal to μ_{cut}. Suppose that after ν shrinking steps the cap prevents chiseling from shrinking the current region any further. It makes sense to set $\alpha_t = 0$ for all $t < \nu$ since at those stages we still estimate that $\mathcal{R}_t$ contains points with negative conditional mean. In Section 3.4, our simulations tested $H_0 : \mu(\mathcal{R}_\nu) \leq \mu_{\text{cut}}$ at the full level α, i.e. $\alpha_\nu = \alpha$. But this may not be optimal, for example, because of estimation uncertainty. A natural way to accommodate multiple comparisons is to allocate α "equally" between stage ν and some future stages.

Consider the following strategy. Initialize $\alpha_{\text{spent}} = 0$ and choose some minimum sample size $n_{\min} > 0$ and initial split proportion $p \in (0, 1)$. If it is necessary to enforce Constraint 3.12, take $\alpha_{\min} > 0$, but let $\alpha_{\min} = 0$ otherwise. Initialize chiseling with a p fraction of the full sample as in Section 3.4 and for $t < \nu$ let $\alpha_t = 0$, while at each stage $t \geq \nu$, (1) set $\alpha_{\text{budget}} \leftarrow \left(\frac{n_\nu - n_t}{n_\nu - n_{\min}}\right) \times \alpha$ if $n_t \geq n_{\min}$ and α otherwise, (2) calculate $\alpha'_t \leftarrow 1 - \frac{1 - \alpha_{\text{budget}}}{1 - \alpha_{\text{spent}}}$ and test at level $\alpha_t \leftarrow \alpha'_t \cdot \mathbb{1}\{\alpha'_t \geq \alpha_{\min}\}$, then (3) update $\alpha_{\text{spent}} \leftarrow 1 - (1 - \alpha_{\text{spent}})(1 - \alpha_t)$. Essentially, this sets the remaining α proportional to the remaining sample size. During these stages, continue refitting the conditional mean $\hat{\mu}(\cdot)$ and shrinking the region using $f(\cdot) = \hat{\mu}(\cdot)$ and $c = \infty$ so that new data is always revealed. Sometimes, it may be reasonable to spend a bit of the error budget on testing $\mathcal{R}_0 = \mathcal{X}$, i.e. let $\alpha_0 > 0$ (see Section 4.3). In this case, initialize $\alpha_{\text{spent}} \leftarrow \alpha_0$ and calculate $\alpha_{\text{budget}} \leftarrow \alpha_0 + \left(\frac{n_\nu - n_t}{n_\nu - n_{\min}}\right) \times (\alpha - \alpha_0)$ for $t \geq \nu$, which generalizes the above.

This strategy is intentionally simple. While more sophisticated α allocation strategies

can lead to more powerful procedures, we find throughout Section 5 that even this simple approach dramatically outperforms data splitting, and moreover improves over the variant of chiseling in Section 3.4 that spends all of α at step ν. We leave an exploration of more sophisticated strategies to future work.

4.3 Subgroup selection when non-responders are not harmed

In some applications including certain clinical RCTs, it may be implausible for the CATE to ever be negative, e.g., since such trials are only approved if there is a strong prior belief that the treatment will not bring harm to any individuals. Though subgroup selection can still be used to discover units with an effect much larger than zero (i.e. by taking $\mu_{\text{cut}} > 0$), it can bring value even when $\mu_{\text{cut}} = 0$ and $\mu(x) \geq 0$ for all $x \in \mathcal{X}$, so that $\mathcal{X}$ is an optimal subgroup. This is because some units may have small or zero treatment effects and thus excluding them from the subgroup can boost the power of a test of significance (and hence the expected utility of the reported region). We will see a numerical demonstration of this in Section 5.5.

When $\mathcal{X}$ is believed to be optimal, however, it is also natural to compare chiseling to the procedure that tests $H_0 : \mu(\mathcal{X}) \leq 0$ using a one-sided t-test, reporting $\mathcal{X}$ if a rejection is made and $\emptyset$ otherwise; we will refer to this procedure as the 'global t-test' for short. The global t-test always tests an optimal region, and when treatment effects are positive and homogeneous, it is essentially an optimal procedure. Luckily, we can instantiate chiseling in such a way that even when effects are homogeneous, chiseling is not much worse than the global t-test, but when a sufficient fraction of units have zero or nearly zero treatment effects (but are still all nonnegative, so $\mathcal{X}$ remains optimal), chiseling outperforms the global t-test. This is possible by tuning α_0; setting $\alpha_0 = 0$ puts no stock in testing $\mathcal{X}$, while setting $\alpha_0 = \alpha$ reduces to just testing $\mathcal{X}$ and is essentially identical to the global t-test procedure.[7] Anything in-between trades off between these extremes, and in Appendix C.13 we show that one can bound the worst case loss in expected utility of chiseling when compared to the t-test as a function of α_0. For instance, if $\mu(\cdot)$ is nonnegative everywhere and $\alpha_0 = \alpha/2$, then the normalized expected utility of chiseling cannot be less than that of the t-test by more than 12.5 percentage points (asymptotically), regardless of how chiseling is otherwise instantiated, but can potentially be much greater (Section 5.5).

4.4 Computational efficiency

While updating $\hat{\mu}(\cdot)$, the current estimate of $\mu(\cdot)$, each time a single new data point is revealed is ideal from the standpoint of statistical efficiency, it may not be practical if the machine learning method is computationally expensive and/or the dataset is large. Luckily, some machine learning methods can be efficiently updated upon observing a single new data point, e.g. rank-one updating ridge regression or online stochastic gradient descent, but since chiseling is interactive it is easy to enforce computational constraints as needed. For instance by only updating $\hat{\mu}(\cdot)$ every $B > 1$ calls to Algorithm 2, or "warm starting" fitting procedures by initializing parameters at previous estimates.

---

[7] At $t = 0$, the asymptotic test we construct in Section 3.6 reduces to running a one-sided Z-test on $\mathcal{X}$, which is asymptotically equivalent to the one-sided t-test.

4.5 Constructing interpretable regions

In practice, investigators may prefer subgroups that are easy to interpret so that it is easier to assess the plausibility of the discovered subgroups, in addition to being easier to communicate to stakeholders. There are many different criteria for what makes a subgroup interpretable, and each different criteria will necessitate a different way of using chiseling to produce a satisfactorily interpretable region. In this subsection, we consider one common standard for interpretability: regions that are axis-aligned hyperrectangles. That is, we desire regions of the form $\mathcal{R} = (a_1, b_1) \times ... \times (a_d, b_d) \subseteq \mathbb{R}^d$, so that checking whether $x \in \mathcal{R}$ reduces to checking whether each of d covariates individually lies in an interval. If desired, a notion of sparsity may be incorporated by encouraging $(a_j, b_j) = (-\infty, \infty)$ for most j.

Let $x^{(j)}$ denote the jth element of a $x \in \mathbb{R}^d$. We have presented chiseling as shrinking along the upper level sets of a function $f(\cdot)$, and in order to ensure the upper level sets of $f(\cdot)$ are hyperrectangular, it suffices to consider the form $f(x) = \min_{j=1,...,d}\{f_j(x^{(j)})\}$, where each $f_j : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ is any function whose upper level sets are intervals (we call such a function "unimodal"). This is because for any $z \in \mathbb{R}$,

$$f(x) > z \iff \forall j \ f_j(x^{(j)}) > z \iff \forall j \ x^{(j)} \in \{x' \in \mathbb{R} : f_j(x') > z\},$$

thus implying that $\{x \in \mathbb{R}^d : f(x) > z\}$ is the product of intervals. Setting $f_j(x^{(j)}) = \infty$ for most j facilitates sparsity. In fact, any $f(\cdot)$ with hyperrectangular upper level sets has this form (Appendix C.14), so this is a fully general formulation. The intersection of hyperrectangles remains a hyperrectangle, so using scoring functions of the above form in each shrinking step will result in a sequence of chiseled regions that are hyperrectangles. The endpoints of any region reported by chiseling are straightforward to calculate by simply keeping track of the $2d$ coordinate endpoints of each hyperrectangle at each stage of Algorithm 1 and intersecting them.

The individual functions $f_j(\cdot)$ need to be learned from the data. A simple idea, showcased in Section 6, is to let $f_j(\cdot)$ be an estimate of $\mathbb{E}[Y \mid X^{(j)}]$ using all of the data available thus far that also enforces the structural requirement of unimodality. Examples of unimodal regressions include isotonic regression and constrained quadratic regression. A further challenge is that in this case $f(\cdot)$ is no longer an estimate of $\hat{\mu}(\cdot)$, and hence there is no special reason to begin testing at the point where $f(\cdot)$ crosses 0. Instead, a naive workaround (which in Section 6 we find to be effective) is to run the strategy outlined in Section 4.2 but with $\nu = 0$ so that α is allocated "equally" over the entire sequence of chiseled regions.

5 Numerical studies

In this section, we verify using numerical experiments that chiseling possesses valid Type I error control, which is exact or nearly exact at the boundary of the global null, and that regions reported by chiseling have higher expected utility than those reported by methods with comparable guarantees. In particular, we consider the following methods:

- *p*-**chiseling**. The version of chiseling described in Section 4.2 with initial split proportion p.

- **p-data splitting.** Data splitting with initial split proportion p as described in Section 3.4.

- **p-simultaneous data splitting.** An approach that leverages data splitting but then produces simultaneously valid tests for a nested collection of regions. This is representative of the methods of Bonetti (2004), Song and Pepe (2004), and Imai and Li (2023), which are all based on a common principle. In particular, this method uses a training set of proportion p to define 10 nested regions where the largest is the estimated optimal region, e.g., the one that p-data splitting would test. It reports the largest rejected region. See Appendix B.2.1 for details.

Because the parameter p may be difficult to tune, where appropriate we will (1) show results as a function of p, and (2) show results for the Bonferroni aggregation (described in Section 4.1) across $p \in \{0.2, 0.5, 0.8\}$, letting the target level be $\alpha/3$ in each of the three instances.

We will also consider the **global t-test** procedure from Section 4.3, an **oracle data splitting** approach which has access to twice the sample size and runs 0.5-data splitting, and an **oracle simultaneous data splitting** approach which has access to twice the sample size and runs 0.5-simultaneous data splitting. Effectively, the oracle methods tell us what utility to expect if we were able to use the entire sample both for subgroup selection and for testing in a valid way, and thus provide approximate upper bounds on the achievable utility.

Throughout this section, we only consider the version of chiseling described in Section 4.2, which requires the multiple testing framework we have developed, as opposed to the version described in Section 3.4, which tests only a single region. While we consistently find that the latter approach dominates data splitting for any given p (Figure 2), we showcase the former because we find that it is uniformly more powerful than the latter (see Appendix B.4).

High-level details. Throughout, n will denote the sample size and p is synonymous with the initial split proportion. We let $\alpha = 0.05$, $\alpha_{\min} = 1 - (1 - \alpha)^{1/40}$, and $n_{\min} = 30$. All expected utilities are normalized by the utility of the optimal region. While maximizing expected utility is our primary goal, we report power curves for every setting in Appendix B.5. Simulations are repeated at least 2,500 times, and all standard errors are less than 1%. All additional details can be found in Appendix B.2.

5.1 Type I error and exactness

Our simulations confirm that, at the boundary of the global null, the Type I error of chiseling is essentially exact across a diverse array of data generating processes and choices of machine learning algorithm within chiseling. We confirm the quality of the asymptotic approximation when the IPW transformation is used to construct the outcomes Y, as well as for when a misspecified AIPW transformation is used. To save space, we relegate figures and additional details to Appendix B.1, but we broadly report that the quality of our asymptotic approximation is high and that the Type I error is essentially exact even when the minimum sample size corresponding to any test is as small as 30. Next, we confirm that the Type I error of chiseling is controlled when we are not under the global null. In each of the simulation settings in the subsequent sections (none of which are under the global null), we also

calculate the Type I error $\mathbb{P}(\mu(\mathcal{R}_\tau) \leq \mu_{\mathrm{cut}})$ and find that the maximum Type I error across any setting is no greater than 1%.

5.2 Binary regression

We consider a binary regression problem where $\mu_{\mathrm{cut}} = 0.9$, $d = 100$, Y follows a logistic regression model in X, and the covariates are mildly correlated. The machine learning algorithm employed by every method is l^2-penalized logistic regression with scikit-learn's default regularization settings, and we vary the probability mass of the optimal subgroup ($\mathbb{P}(X \in \mathcal{R}^*)$ where $\mathcal{R}^*$ is the optimal subgroup) from 50% to 1% and tune n so that the expected utility is nontrivial. Our findings are displayed in Figure 3. Because in this case $\mathbb{P}(Y = 1) < 0.9$ in every setting, we do not report results for the global t-test (its expected utility is negative). For the data splitting methods, an exact binomial test is used in place of a t-test. We find that the expected utility of chiseling dominates that of the other methods across essentially all initial split proportions and optimal subgroup probability masses, and that the Bonferroni aggregated variant of chiseling dominates the Bonferroni aggregated variants of data splitting and simultaneous data splitting. Comparing the best performance of each non-oracle method across initial split proportions, we see that the peak expected utility for chiseling exceeds that of the next-best method by 40–150%. In some situations, chiseling even outperforms the oracles which utilize twice the sample size, which is remarkable but by no means impossible since chiseling leverages a multiple testing framework that is different than that used by the oracles.
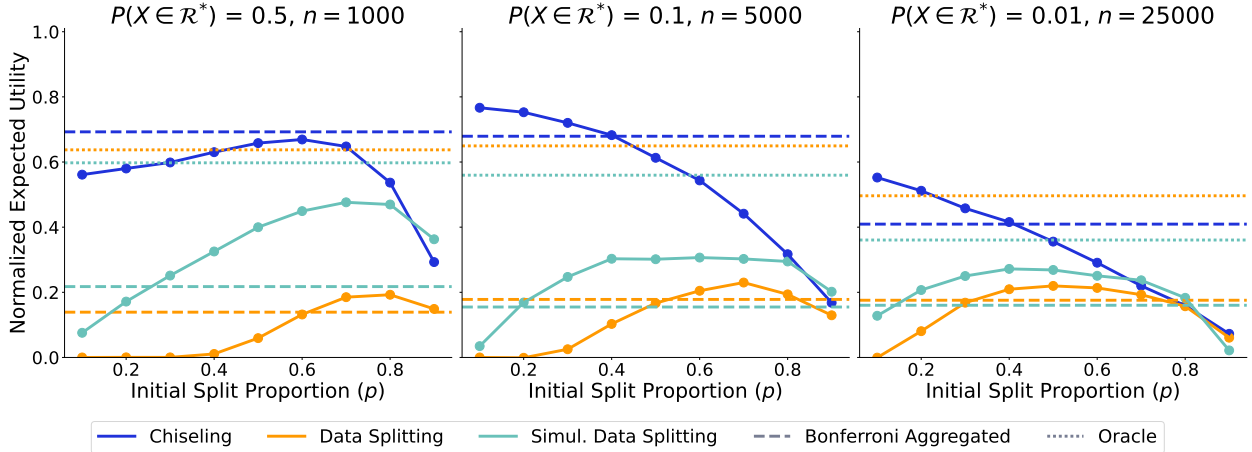

Figure 3: Normalized expected utility in a simulated binary regression problem as the probability mass of the optimal subgroup varies from 50% to 1%. Solid horizontal lines correspond to using Bonferroni to aggregate across initial split proportions $\in \{0.2, 0.5, 0.8\}$ for the method with the corresponding color. The machine learning method used is ℓ^2-penalized logistic regression.

5.3 Heterogeneous randomized control trial

We consider a simulated RCT where $\mu_{\mathrm{cut}} = 0$, $d = 100$, the CATE is linear in X, and the covariates are mildly correlated. The treatments are assigned using a fair coin flip, and we

apply the IPW transformation to convert each instance to a regression problem. The machine learning algorithm we use here is the lasso where the penalty is selected via cross-validation. Like in the previous subsection, we vary the probability mass of the optimal subgroup from 50% to 1%, and we do not report results for the global t-test because the full population $\mathcal{X}$ is under the null. We find that the expected utility of chiseling dominates that of the other methods across all initial split proportions and optimal subgroup probability masses, and that the Bonferroni aggregated variant of chiseling dominates the Bonferroni aggregated variants of data splitting and simultaneous data splitting (Figure 4). Comparing the best performance of each method across initial split proportions, we see that the peak expected utility for chiseling exceeds that of the next-best method by 20–59%.
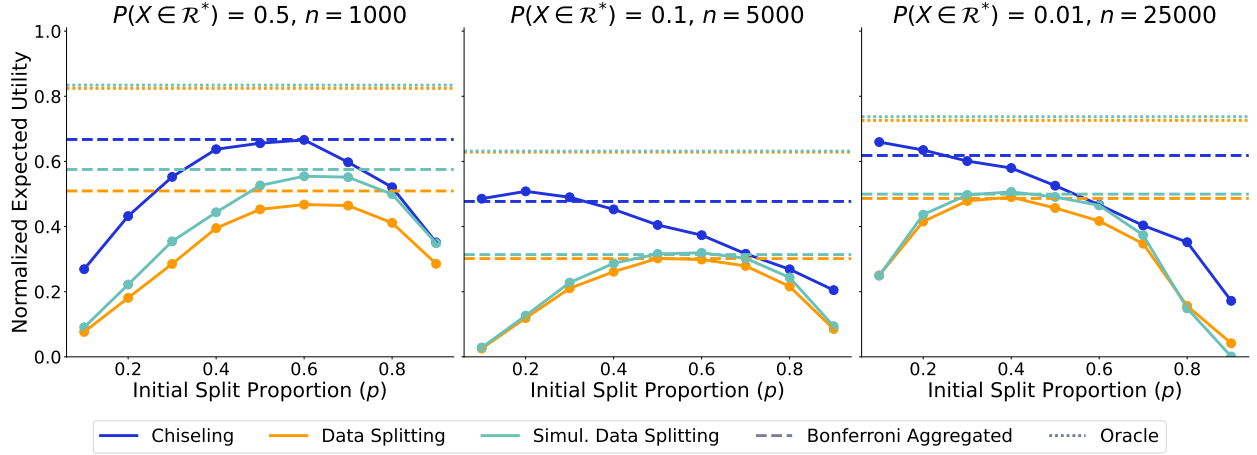

Figure 4: Normalized expected utility in a simulated heterogeneous RCT as the probability mass of the optimal subgroup varies from 50% to 1%. Solid horizontal lines correspond to using Bonferroni to aggregate across initial split proportions $\in \{0.2, 0.5, 0.8\}$ for the method with the corresponding color. The machine learning method used is cross-validated lasso regression.

5.4 RCT with misspecified machine learning model

We empirically investigate how misspecification of the machine learning model affects the relative performances of the subgroup selection procedures. Recall that the methods we consider are valid despite the misspecification. We let the CATE be non-linear in X but use linear regression with a ridge penalty selected using cross-validation as the base machine learning method. Across a variety of optimal subgroup probability masses ranging from 75% to 25%, we find that the expected utility of chiseling dominates that of other methods, including when results are aggregated using Bonferroni. As the expected utility curves do not look substantially qualitatively different from those seen thus far, we relegate the plots to Appendix B.3 (Figures B2 and B12). In these simulations, the peak expected utility for chiseling exceeds that of the next-best method by 13–40%.

22

5.5 RCT with no or small negative treatment effects

We consider a simulated clinical RCT where the treatment does not bring substantial harm to any individuals. Moreover, $\mu_{\mathrm{cut}} = 0$, so that we wish to include units with any positive CATE while being agnostic about units with zero CATE.

First, suppose that the CATE only ever takes on values in $\{0, \tau\}$ for some constant τ so that $\mathcal{X}$ is an optimal subgroup. We choose the CATE to be linear in $X \in \mathbb{R}^5$ and restrict the support of X so that the CATE lies in $\{0, \tau\}$ (see Appendix B.2 for details; we choose this stylized setting to make the distribution of the CATEs maximally interpretable). Chiseling uses linear regression. We vary q, the proportion of units with CATE equal to τ. As per Section 4.3, we let $\alpha_0 = \alpha/2$ and $p = 0.1$ so that asymptotically the normalized expected utility of chiseling can not be more than 12.5 percentage points smaller than that of the global t-test for any q or τ. The left panel of Figure 5 empirically confirms this at the extreme where $q = 1$. But when the proportion of responders is $q = 0.1$, chiseling doubles the expected utility of the global t-test. Also, the point at which it becomes more beneficial to use chiseling is roughly the same as the point at which the oracle methods outperform the global t-test, suggesting that chiseling only struggles when subgroup selection is intrinsically difficult.

The second panel in Figure 5 shows the same as the above, but where the CATE takes on values in $\{-0.1\tau, \tau\}$; that is, the treatment is 10 times more effective for responders than it is harmful for non-responders. It paints the same picture but naturally yields greater potential gains over the global t-test. For instance, when $q = 10\%$, chiseling achieves a normalized expected utility of over 70% while the global t-test has zero expected utility (since $\mu(\mathcal{X}) = 0$).
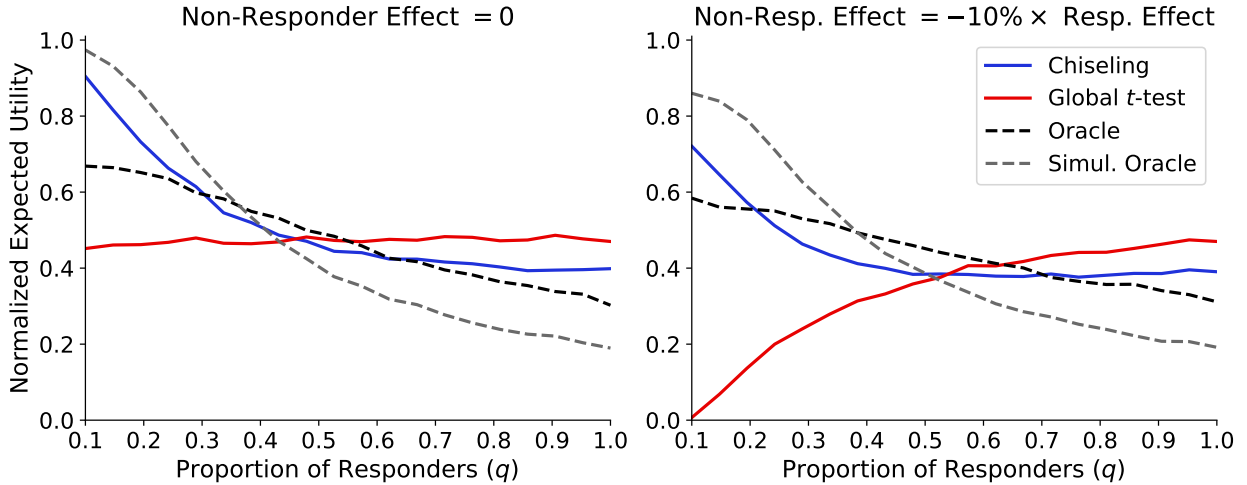

Figure 5: Normalized expected utility in a simulated RCT where the treatment effect for responders is τ, and where non-responders either have 0 treatment effect (left panel) or treatment effect equal to -0.1τ (right panel). Sample size is $n = 1000$ and $\tau = 0.1/q$ where q is the proportion of responders. The machine learning method used is linear regression. We initialize chiseling with $p = 0.1$ and $\alpha_0 = \alpha/2$.
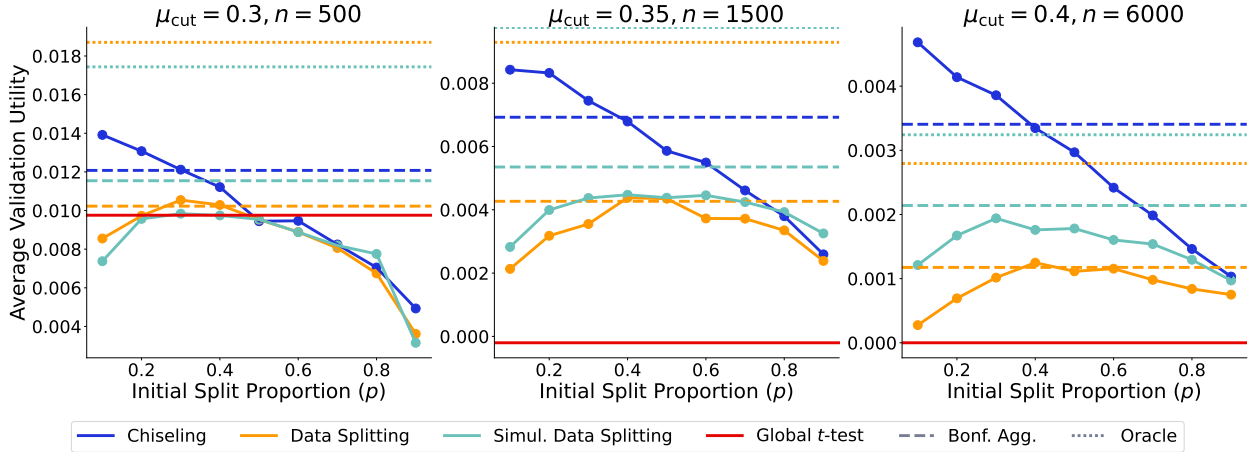
6 Real data application

Figure 6: Average validation utilities on the GSS data across different testing cutoffs μ_{cut}.

In the mid-1980s, the General Social Survey (GSS) incorporated an experiment into its survey of American adults in order to assess whether the wording of a question could influence expressed support for government spending. Survey takers were randomized into either receiving a question which described the spending program as "welfare" or one which described the spending program as "assistance to the poor." Many analyses have replicated the famous ≈ 0.35 increase in probability of support for the spending program when the wording is changed from "welfare" to "assistance to the poor." The experiment was reanalyzed in Green and Kern (2012) using Bayesian Additive Regression Trees in order to understand the heterogeneity in the effect across different demographic characteristics and survey responses.

Imitating the analysis of Green and Kern (2012), we take 36,501 survey data points from the GSS and apply minimal preprocessing (impute non-responses, drop missing data, and one-hot encode non-ordinal categorical variables). This produces $d = 152$ features. We then test the ability of methods to detect subgroups with effects greater than $\mu_{\text{cut}} \in \{0.3, 0.35, 0.4\}$. In order to assess the average utility of the reported subgroups, we repeatedly split the dataset into a training set which is used to run the subgroup selection methods and a validation set which is used to estimate the utilities of the reported subgroups. Note that for each method, both selection and testing are performed strictly on the training set. We appropriately set the size of the training set, which we refer to as n, for each setting so that the expected utility is nontrivial. We apply a simple AIPW transformation with only an intercept and no covariates, and we use the T-learner (Künzel et al., 2019) with scikit-learn's default implementation of the random forest classifier as the base learner.

We see in Figure 6 that the average utility of chiseling dominates that of the other methods in all settings (results for power and expected probability mass may be found in Appendix B.3). Comparing the best performance of each method across initial split proportions, we see that the peak expected utility for chiseling exceeds that of the next-best method by 32–141%. When $\mu_{\text{cut}} = 0.4$, chiseling outperforms the oracle methods, though we caution that the repeated train/validation splits are not independent and hence we cannot produce reliable standard errors for these plots when n is a sizable fraction of 36,501.

Next, we run chiseling on the full 36,501 data points with $\mu_{\mathrm{cut}} = 0.45$. To construct an interpretable subgroup, we use the approach described in Section 4.5 with coordinate-wise isotonic regression (direction adaptively chosen using Spearman's rank correlation) to discover hyperrectangular regions. We initialize chiseling with 1,000 randomly revealed data points. Chiseling discovers a subgroup that includes roughly 4% of the sample. In this subgroup, the empirical difference between treatment and control means is 0.50. We examine in Appendix B.3 (Table 1) the ten most important features as measured by exclusivity of that feature's range in the discovered subgroup, and find strong agreement with previous analyses such as Green and Kern (2012). Further investigation suggests these ten features are the dominant drivers of the effect size in the discovered subgroup; see Figure B5 for details.

7   Discussion

We presented a novel method for subgroup selection that maintains the leanness of data splitting's assumptions but is able to utilize the given sample much more efficiently. It consists of two primary methodological insights: that a specific way of shrinking a subgroup leaves it untarnished, and that we can exactly (or essentially exactly) sequentially test multiple regions while shrinking. In its most basic form, chiseling information-theoretically dominates data splitting for subgroup selection, and we showed that these gains are realized empirically via natural instantiations of chiseling.

The chiseling framework is extremely flexible, and while this paper provided some guidance and default choices for its use, we ultimately believe its flexibility is its key strength, allowing users to incorporate any and all qualitative and quantitative knowledge however they want without having to worry about validity. We list here a few directions for future research on leveraging and further expanding this flexibility.

- *Can we better shrink the regions?* We recommended shrinking regions along an estimate $\hat{\mu}(\cdot)$ of $\mu(\cdot)$, but more sophisticated ideas could be superior. For instance, Spiess et al. (2023) suggest that power can be gained by considering how the variance of the outcome varies across the covariate space, favoring regions with moderate effects and small variances to those with large effects and outsized variances.

- *Can we better allocate α?* We suggested a simple strategy for allocating α, but it may not be the most powerful choice. Relatedly, how should one set α when $\hat{\mu}(\cdot)$ is grossly misspecified? While $\hat{\mu}(\cdot)$ may still be effective for shrinking, the value of $\hat{\mu}(\cdot)$ cannot be trusted to indicate the optimal region threshold. Ideas from the calibration literature could prove useful.

- *Can we better initialize chiseling?* Can the initial split proportion p be chosen adaptively, or even done away with entirely? If prior information exists, how can it be effectively incorporated into the initialization? For an idea along these lines, see Appendix C.11 for an initialization inspired by the "two team cross-screening" approach of Roy et al. (2025).

- *How should chiseling be used to maximize other utility functions?* The right way to use chiseling will differ based on the utility one wishes to maximize, and in particular may be different if one wishes to maximize power or expected probability mass.

- *Can we give inferences for other estimands?* The structure of the naive mean estimator yields tractable conditioning events that we take advantage of in our sequential testing framework. We also showed that one can utilize sophisticated transformations such as the AIPW transformation with known propensities, but additional work is required to extend our approach to the observational setting. It would be interesting to see if estimands that are not based on an average of transformed outcomes (e.g. conditional medians or M-estimands) can be addressed in our framework.

- *Can we broaden the inferential targets?* Chiseling, as we have developed in this paper, is a method for rejecting one region at a fixed cutoff, but there may be situations where an analyst wishes to reject multiple regions, or to pair the selected subgroup with a point estimate or tight confidence interval for its effect. Though we discuss some multiple testing extensions in Appendix C.15 and methods for point estimates and simultaneous confidence intervals in Appendix C.16, there is more work to be done.

- *Can we expand the action space of chiseling?* Currently, there are two basic actions one can take at any point during chiseling: reveal a random point from the current region (see Appendix C.4) or shrink the region. What other actions can be added to this set? For instance, are there techniques that would allow one to restore erroneously excised regions of the covariate space? Instead of revealing a point uniformly at random, can we reveal it according to a weight function? And can we reveal points based not only on X but also on Y while still permitting valid inference?

Acknowledgments

References

Andrews, I. and Chen, J. (2025). Certified Decisions. arXiv: 2502.17830 [econ.EM].

Andrews, I., Kitagawa, T., and McCloskey, A. (2023). Inference on Winners. *The Quarterly Journal of Economics* 139.1, pp. 305–358.

Armitage, P., McPherson, C. K., and Rowe, B. C. (1969). Repeated Significance Tests on Accumulating Data. *Journal of the Royal Statistical Society. Series A (General)* 132.2, p. 235.

Armstrong, T. B. and Shen, S. (2015). Inference on optimal treatment assignments. *SSRN Electronic Journal*.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113.27, pp. 7353–7360.

Athey, S. and Wager, S. (2021). Policy Learning With Observational Data. *Econometrica* 89.1, pp. 133–161.

Austern, M. and Zhou, W. (2020). Asymptotics of Cross-Validation. arXiv: 2001.11111 [math.ST].

Bahr, B. von and Esseen, C.-G. (1965). Inequalities for the rth Absolute Moment of a Sum of Random Variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics* 36.1, pp. 299–303.

Baricz, Á. (2008). Mills' ratio: Monotonicity patterns and functional inequalities. *Journal of Mathematical Analysis and Applications* 340.2, pp. 1362–1370.

Bartroff, J., Lai, T. L., and Shih, M.-C. (2013). *Sequential Experimentation in Clinical Trials: Design and Analysis*. Springer New York.

Bates, S., Hastie, T., and Tibshirani, R. (2023). Cross-Validation: What Does It Estimate and How Well Does It Do It? *Journal of the American Statistical Association* 119.546, pp. 1434–1445.

Bayle, A., Janson, L., and Mackey, L. (2025). The Relative Instability of Model Comparison with Cross-validation. arXiv: 2508.04409 [stat.ML].

Bayle, P. et al. (2020). Cross-validation Confidence Intervals for Test Error. *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 16339–16350.

Benjamini, Y. and Hochberg, Y. (2000). On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics. *Journal of Educational and Behavioral Statistics* 25.1, p. 60.

Berger, J. O., Wang, X., and Shen, L. (2014). A Bayesian approach to subgroup identification. *Journal of Biopharmaceutical Statistics* 24.1, pp. 110–129.

Bertsimas, D., Korolko, N., and Weinstein, A. M. (2019). Identifying Exceptional Responders in Randomized Trials: An Optimization Approach. *INFORMS Journal on Optimization* 1.3, pp. 187–199.

Bonetti, M. (2004). Patterns of treatment effects in subsets of patients in clinical trials. *Biostatistics* 5.3, pp. 465–481.

Cai, H. et al. (2022). CAPITAL: Optimal subgroup identification via constrained policy tree search. *Statistics in Medicine* 41.21, pp. 4227–4244.

Cai, T. et al. (2010). Calibrating parametric subject-specific risk estimation. *Biometrika* 97.2, pp. 389–404.

Cai, T. et al. (2011). Analysis of randomized comparative clinical trial data for personalized treatment selections. *Biostatistics* 12.2, pp. 270–282.

Chao, P. and Fithian, W. (2021). AdaPT-GMM: Powerful and robust covariate-assisted multiple testing. arXiv: 2106.15812 [stat.ME].

Chernozhukov, V. et al. (2018). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India. Tech. rep. Cambridge, MA: National Bureau of Economic Research.

Chernozhukov, V. et al. (2025). Policy Learning with Confidence. arXiv: 2502.10653 [econ.EM].

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4.1.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika* 62.2, pp. 441–444.

Demets, D. L. and Lan, K. K. G. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine* 13.13–14, pp. 1341–1352.

Ding, Y., Lin, H.-M., and Hsu, J. C. (2016). Subgroup mixable inference on treatment efficacy in mixture populations, with an application to time-to-event outcomes. *Statistics in Medicine* 35.10, pp. 1580–1594.

Dixon, D. O. and Simon, R. (1991). Bayesian Subset Analysis. *Biometrics* 47.3, p. 871.

Duan, B., Ramdas, A., and Wasserman, L. (2020a). Familywise Error Rate Control by Interactive Unmasking. *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 2720–2729.

— (2022). Interactive rank testing by betting. *Proceedings of the First Conference on Causal Learning and Reasoning*. Ed. by B. Schölkopf, C. Uhler, and K. Zhang. Vol. 177. Proceedings of Machine Learning Research. PMLR, pp. 201–235.

Duan, B., Wasserman, L., and Ramdas, A. (2024). Interactive identification of individuals with positive treatment effect while controlling false discoveries. *Journal of Causal Inference* 12.1.

Duan, B. et al. (2020b). Interactive martingale tests for the global null. *Electronic Journal of Statistics* 14.2.

Fan, Q. et al. (2020). Estimation of Conditional Average Treatment Effects With High-Dimensional Data. *Journal of Business & Economic Statistics* 40.1, pp. 313–327.

Fithian, W., Sun, D., and Taylor, J. (2017). Optimal Inference After Model Selection. arXiv: 1410.2597 [math.ST].

Foster, D. P. and Stine, R. A. (2008). α-Investing: a Procedure for Sequential Control of Expected False Discoveries. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70.2, pp. 429–444.

Foster, J. C., Taylor, J. M. G., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 30.24, pp. 2867–2880.

Fraser, D. A. S. (1951). Sequentially Determined Statistically Equivalent Blocks. *The Annals of Mathematical Statistics* 22.3, pp. 372–381.

Friedman, J. H. and Fisher, N. I. (1999). Bump hunting in high-dimensional data. *Statistics and Computing* 9.2, pp. 123–143.

G'Sell, M. G. et al. (2015). Sequential Selection Procedures and False Discovery Rate Control. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 78.2, pp. 423–444.

Green, D. P. and Kern, H. L. (2012). Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees. *Public Opinion Quarterly* 76.3, pp. 491–511.

Guo, X. and He, X. (2020). Inference on Selected Subgroups in Clinical Trials. *Journal of the American Statistical Association* 116.535, pp. 1498–1506.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects (with Discussion). *Bayesian Analysis* 15.3.

Haybittle, J. L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *The British Journal of Radiology* 44.526, pp. 793–797.

Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240.

Hirano, K. and Porter, J. R. (2009). Asymptotics for Statistical Treatment Rules. *Econometrica* 77.5, pp. 1683–1701.

Huang, M., Tang, T. M., and Kenney, A. M. (2025). Distilling heterogeneous treatment effects: Stable subgroup estimation in causal inference. arXiv: 2502.07275 [stat.ME].

Imai, K. and Li, M. L. (2023). Experimental evaluation of individualized treatment rules. *Journal of the American Statistical Association* 118.541, pp. 242–256.

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics* 7.1, pp. 443–470.

Javanmard, A. and Montanari, A. (2015). On Online Control of False Discovery Rate. arXiv: 1502.06197 [stat.ME].

Jia, Z., Imai, K., and Li, M. L. (2024). The Cram Method for Efficient Simultaneous Learning and Evaluation. arXiv: 2403.07031 [cs.LG].

Jin, Y. and Candes, E. J. (2023). Selection by Prediction with Conformal p-values. *Journal of Machine Learning Research* 24.244, pp. 1–41.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science* 22.4.

Kehl, V. and Ulm, K. (2006). Responder identification in clinical trials with censored data. *Computational Statistics & Data Analysis* 50.5, pp. 1338–1355.

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics* 17.2.

Kitagawa, T. and Tetenov, A. (2018). Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* 86.2, pp. 591–616.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer New York.

Kovalchik, S. A., Varadhan, R., and Weiss, C. O. (2013). Assessing heterogeneity of treatment effect in a clinical trial with the proportional interactions model. *Statistics in Medicine* 32.28, pp. 4906–4923.

Künzel, S. R. et al. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116.10, pp. 4156–4165.

Ladhania, R. et al. (2020). Learning and Testing Sub-groups with Heterogeneous Treatment Effects:A Sequence of Two Studies. arXiv: 2006.11628 [stat.ME].

Lai, T. L., Lavori, P. W., and Liao, O. Y.-W. (2014). Adaptive choice of patient subgroup for comparing two treatments. *Contemporary Clinical Trials* 39.2, pp. 191–200.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete Sequential Boundaries for Clinical Trials. *Biometrika* 70.3, p. 659.

Lei, L. and Candès, E. J. (2021). Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 83.5, pp. 911–938.

Lei, L. and Fithian, W. (2018). AdaPT: An Interactive Procedure for Multiple Testing with Side Information. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.4, pp. 649–679.

Lei, L., Ramdas, A., and Fithian, W. (2020). A general interactive framework for false discovery rate control under structural constraints. *Biometrika* 108.2, pp. 253–267.

Liang, R. and Barber, R. F. (2025). Algorithmic stability implies training-conditional coverage for distribution-free prediction methods. *The Annals of Statistics*.

Lipkovich, I. and Dmitrienko, A. (2014). Strategies for Identifying Predictive Biomarkers and Subgroups with Enhanced Treatment Effect in Clinical Trials Using SIDES. *Journal of Biopharmaceutical Statistics* 24.1, pp. 130–153.

Lipkovich, I. et al. (2011). Subgroup identification based on differential effect search—A recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine* 30.21, pp. 2601–2621.

Luedtke, A. and Chambaz, A. (2020). Performance guarantees for policy learning. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 56.3, pp. 2162–2188.

Luedtke, A. R. and Laan, M. J. van der (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *The Annals of Statistics* 44.2, pp. 713–742.

Ma, S. and Huang, J. (2017). A Concave Pairwise Fusion Approach to Subgroup Analysis. *Journal of the American Statistical Association* 112.517, pp. 410–423.

Manski, C. F. (2004). Statistical treatment rules for heterogeneous populations. *Econometrica* 72.4, pp. 1221–1246.

Medarametla, D. and Candès, E. (2021). Distribution-free conditional median inference. *Electronic Journal of Statistics* 15.2.

Müller, M. M. et al. (2024). Isotonic subgroup selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 87.1, pp. 132–156.

Nie, X. and Wager, S. (2020). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108.2, pp. 299–319.

O'Brien, P. C. and Fleming, T. R. (1979). A Multiple Testing Procedure for Clinical Trials. *Biometrics* 35.3, p. 549.

Oganisian, A., Mitra, N., and Roy, J. (2020). Bayesian Nonparametric Cost-Effectiveness Analyses: Causal Estimation and Adaptive Subgroup Discovery. arXiv: 2002.04706 [stat.ME].

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* 64.2, pp. 191–199.

Ponomarev, K. and Semenova, V. (2024). On the Lower Confidence Band for the Optimal Welfare. arXiv: 2410.07443 [econ.EM].

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics* 39.2.

Reeve, H. W. J., Cannings, T. I., and Samworth, R. J. (2023). Optimal subgroup selection. *The Annals of Statistics* 51.6.

Ritzwoller, D. M. and Syrgkanis, V. (2024). Simultaneous Inference for Local Structural Parameters with Random Forests. arXiv: 2405.07860 [econ.EM].

Robertson, D. S., Wason, J. M. S., and Ramdas, A. (2023). Online Multiple Hypothesis Testing. *Statistical Science* 38.4.

Roy, S. et al. (2025). Exploration, Confirmation, and Replication in the Same Observational Study: A Two Team Cross-Screening Approach to Studying the Effect of Unwanted Pregnancy on Mothers' Later Life Outcomes. arXiv: 2505.14480 [stat.ME].

Schnell, P. et al. (2017). Subgroup inference for multiple treatments and multiple endpoints in an Alzheimer's disease treatment trial. *The Annals of Applied Statistics* 11.2, pp. 949–966.

Schnell, P. M. et al. (2016). A Bayesian Credible Subgroups Approach to Identifying Patient Subgroups with Positive Treatment Effects. *Biometrics* 72.4, pp. 1026–1036.

Semenova, V. and Chernozhukov, V. (2020). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal* 24.2, pp. 264–289.

Sivaganesan, S., Laud, P. W., and Müller, P. (2010). A Bayesian subgroup analysis with a zero-enriched Polya Urn scheme. *Statistics in Medicine* 30.4, pp. 312–323.

Song, X. and Pepe, M. S. (2004). Evaluating Markers for Selecting a Patient's Treatment. *Biometrics* 60.4, pp. 874–883.

Spiess, J., Syrgkanis, V., and Wang, V. Y. (2023). Finding Subgroups with Significant Treatment Effects. arXiv: 2103.07066 [econ.EM].

Storey, J. D. (2002). A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 64.3, pp. 479–498.

Stoye, J. (2009). Minimax regret treatment choice with finite samples. *Journal of Econometrics* 151.1, pp. 70–81.

Su, X. et al. (2009). Subgroup Analysis via Recursive Partitioning. *Journal of Machine Learning Research* 10.5, pp. 141–158.

Talisa, V. B. and Chang, C.-C. H. (2021). Learning and confirming a class of treatment responders in clinical trials. *Statistics in Medicine* 40.22, pp. 4872–4889.

Tian, J. and Ramdas, A. (2021). Online control of the familywise error rate. *Statistical Methods in Medical Research* 30.4, pp. 976–993.

VanderWeele, T. J. et al. (2019). Selecting optimal subgroups for treatment using many covariates. *Epidemiology* 30.3, pp. 334–341.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer-Verlag.

Wager, S. (2024a). Causal Inference: A Statistical Learning Approach. Draft book manuscript, Stanford University, 6 Sep 2024.

— (2024b). Sequential Validation of Treatment Heterogeneity. arXiv: 2405.05534 [econ.EM].

Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association* 113.523, pp. 1228–1242.

Wald, A. (1945). Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics* 16.2, pp. 117–186.

Wan, F., Liu, W., and Bretz, F. (2024). Confidence sets for a level set in linear regression. *Statistics in Medicine* 43.6, pp. 1103–1118.

Yang, S. et al. (2021). Propensity score weighting for causal subgroup analysis. *Statistics in Medicine* 40.19, pp. 4294–4309.

Zhang, B. et al. (2012). A Robust Method for Estimating Optimal Treatment Regimes. *Biometrics* 68.4, pp. 1010–1018.

Zhang, C. and Zheng, Z. (2023). Inference of subgroup-level treatment effects via generic causal tree in observational studies. *JUSTC* 53.11, p. 1102.

Zhang, W. and Misra, S. (2024). Coarse Personalization. *Proceedings of the 25th ACM Conference on Economics and Computation*. EC '24. ACM, pp. 1206–1208.

Zhang, Y. et al. (2021). Subgroup causal effect identification and estimation via matching tree. *Computational Statistics & Data Analysis* 159, p. 107188.

Zhao, B., Ivanova, A., and Fine, J. (2023). Inference on subgroups identified based on a heterogeneous treatment effect in a post hoc analysis of a clinical trial. *Clinical Trials* 20.4, pp. 370–379.

Zhao, Y. et al. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* 107.449, pp. 1106–1118.

Zimmert, M. and Lechner, M. (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. arXiv: 1908.08779 [econ.EM].

Appendix Contents

Appendix A   Extended literature review

A.1   Subgroup learning

There is a vast literature on works that seek to learn subgroups from the data, but which do not by themselves propose ways of evaluating that subgroup without data splitting, or otherwise only provide weak guarantees on the learned subgroup. This includes the literature on policy learning, which is distinguished enough that we discuss it separately in the next subsection. While these works do not focus on the inferential criteria that we consider here, essentially any one of them has potentially important insight to offer about one of the key tuning parameters in chiseling: how to shrink the region.

In this space, tree-based methods have been enormously popular since they naturally form partitions of the covariate space into relatively interpretable subgroups (Su et al., 2009; Lipkovich et al., 2011; Lipkovich and Dmitrienko, 2014; Athey and Imbens, 2016; Ladhania et al., 2020; Zhang et al., 2021; Yang et al., 2021; Cai et al., 2022; Zhang and Zheng, 2023; Huang et al., 2025). The aspects of subgroup learning covered in these works include: splitting criteria, for instance, distinguishing the goals of minimizing the generalization mean-squared error (MSE) and the expected empirical MSE of "honest" estimates calculated on a held out test set; resampling based heuristics for reducing overfitting; variable screening steps; the use of more efficient (doubly robust) estimators of the treatment effect to inform splits; and the incorporation of constraints on subgroup structure. Non tree-based procedures include "bump hunting" (Friedman and Fisher, 1999; Kehl and Ulm, 2006), covariate coarsening via optimal transport (Zhang and Misra, 2024), and mixture modeling (Ding et al., 2016). Furthermore, some works such as Foster et al. (2011), Bertsimas et al. (2019), and Zhao et al. (2023) validate their approaches using informal principles and simulations but are not accompanied by rigorous inferential guarantees.

Lastly, we highlight two works where a subgroup is designed with the specific goal of optimizing the power to replicate a significant subgroup finding on an external dataset. Talisa and Chang (2021) evaluate a subgroup selection procedure favorably if the selected subgroup yields a significant finding on a holdout set. Spiess et al. (2023) initiate a formal study of this problem, and make the interesting observation that the best powered subgroup may not necessarily be the one that maximizes the policy value. This is because certain regions of the covariate space with strong average effects may be offset by high treatment effect variance,

which can deteriorate the power of a significance test. Under a local asymptotic scaling, they write down a formal objective that they develop methods to optimize for. It is also noteworthy that they consider, but do not investigate, a possible extension of their method that, rather than optimizing power, optimizes an expected utility very similar to what we have considered in this work.

A.2 Policy learning

The policy learning literature is broadly distinguished by a focus on regret minimization in a decision-theoretic framework. A now-classical goal in policy learning is to minimize

$$\text{regret} := \mathcal{V}(\pi^*) - \mathbb{E}[\mathcal{V}(\hat{\pi}_n)],$$

where $\mathcal{V}(\pi) := \mathbb{E}[Y(\pi(X))]$, π^* is a maximizer of $\mathcal{V}(\pi)$ over a class of policies Π, and $\hat{\pi}_n$ is the (random, data-dependent) policy that is obtained from a random sample of n data points. Manski (2004) calls $\hat{\pi}_n$ a *statistical treatment rule* (STR). Generally, an STR is judged favorably relative to a class of distributions on $(Y(1), Y(0), W, X)$ if it is *admissible* and *minimax* with regards to regret minimization. In the literature, particular focus has been devoted to establishing the minimaxity of various STRs. Manski (2004) established regret bounds over a restricted set Π, which were subsequently refined and generalized in a number of works (Hirano and Porter, 2009; Stoye, 2009; Kitagawa and Tetenov, 2018). Often, regret bounds are provided in terms of complexity measures of the policy class Π such as the Vapnik–Chervonenkis dimension. An important insight that arises in this literature is that for regret minimization under misspecified parametric models, one should base decisions on direct maximization of the empirical welfare rather than using the plug-in policy that thresholds an estimate of $\mathbb{E}[Y(1) - Y(0) \mid X]$, since the latter may not converge to the welfare maximizing policy due to the misspecification. This is an aspect that we have not explored in our work, but which may offer insights about how to best leverage chiseling.

Additional work in this vein includes the following. Qian and Murphy (2011) use a very large basis expansion of the covariate space to linearly model the conditional mean function $\mathbb{E}[Y \mid X, T]$. Zhao et al. (2012) convert the policy maximization problem into a weighted classification problem where each unit is weighted according to its outcome. Imai and Ratkovic (2013) leverages the support vector machine to produce a sparse, parsimonious model for treatment effect heterogeneity across treatments and subgroups. VanderWeele et al. (2019) describe the population optimal treatment allocation rules under various constraints such as a constraint on the treated proportion or cost/side effect constraints. Luedtke and Chambaz (2020) show that the regret of empirical risk minimization can decay at rates faster than $n^{-1/2}$ for fixed data generating distributions. Athey and Wager (2021) leverage ideas from semi-parametric efficiency theory, using estimates of doubly robust "scores" as pseudo-outcomes and plugging these into an empirical risk minimization framework. We also refer the reader to the "Related Work" section of Athey and Wager (2021) for a survey of developments along the theme of regret minimization.

Note that none of the above discuss inference. Indeed, the ethos of a decision-theoretic solution is generally to obviate a direct need for inference by focusing on the expected gain or cost of each action. There are, however, some exceptions. For instance, some papers do consider inference for the value of the optimal policy or the policy at hand (Zhang et

al., 2012; Luedtke and Laan, 2016; Ponomarev and Semenova, 2024), usually either with the requirement that the policy is learned efficiently or via data splitting. Works such as Chernozhukov et al. (2025) and Andrews and Chen (2025) are still situated in a decision-theoretic framework but discuss the role of uncertainty in decision making: the former absorbs the uncertainty quantification into the objective by utilizing losses that reflect risk aversion, while the latter characterizes the role of confidence sets in decision making.

A.3 Sequential, online, and adaptive hypothesis testing

Sequential hypothesis testing was introduced in Wald (1945) to address the challenge of assessing significance in an experiment that continues making observations until a significant conclusion is reached. For testing a simple null against a simple alternative, Wald (1945) introduced the sequential probability ratio test, which is optimal in the sense of minimizing the expected number of observations needed to yield a significant conclusion. In clinical trials, Wald's work inspired new protocols for repeatedly assessing evidence for treatment efficacy throughout an experiment while controlling Type I error (Armitage et al., 1969; Haybittle, 1971; Pocock, 1977; O'Brien and Fleming, 1979; Lan and DeMets, 1983; Demets and Lan, 1994). At a high level, these methods work by comparing an accumulating test statistic to different rejection boundaries at different time points. In particular, Lan and DeMets (1983) set testing boundaries according to an object they call the α-spending function, which dictates how much of the error budget should have been utilized by each time point. We refer the reader to the introduction of Bartroff et al. (2013) for a history of the field.

Generally, sequential hypothesis testing considers the calibration of accumulating evidence for a single, fixed hypothesis. In online hypothesis testing, the goal is to make a determination about whether to accept or reject different hypotheses based on evidence that arrives over time. The goal might be to control a notion of error such as the family-wise error or the false discovery rate (FDR). This paradigm dates as far back as Foster and Stine (2008), who developed a method for controlling a quantity known as the *marginal false discovery rate* uniformly over all finite stopping times. Subsequently, more powerful methods were developed, and the framework was generalized to accommodate different error rates such as the classical false discovery rate and the FWER (G'Sell et al., 2015; Javanmard and Montanari, 2015; Tian and Ramdas, 2021). We refer the reader to Robertson et al. (2023) for a survey of recent developments.

The label *adaptive* is somewhat loosely used to describe any method that selects certain hyperparameters of a procedure—especially those that are traditionally fixed *a priori*—in a data-dependent way. For instance, some FDR-controlling methods boost power by estimating the true proportion of nulls from the p-values themselves (Benjamini and Hochberg, 2000; Storey, 2002). An *interactive* testing procedure is a special kind of adaptive procedure, but where the adaptivity is driven by a human analyst who interacts with the data—often with arbitrary flexibility—as opposed to a pre-specified adaptive strategy. We have already surveyed interactive hypothesis testing in Section 1.3 and described its relationship to our work. We re-emphasize that while our work is similar in spirit to existing interactive approaches, it addresses a different problem which requires substantially different technical tools to solve. For instance, the proofs of our main ideas do not rely on properties of martingales, while—with the exception of Duan et al. (2020a)—all of the interactive methods we are aware of rely

on some special property of martingales such as the optional stopping theorem or Ville's inequality. Duan et al. (2020a) prove validity of their FWER-controlling procedure by directly characterizing the distribution of false discoveries when the null p-values are independent.

Lastly, we consider some works that are conceptually related but which do not fit neatly into the above categories. Lai et al. (2014) consider the goal of rejecting the null hypothesis of no treatment effect for the full population if possible, and otherwise selecting a subgroup (from a fixed, *a priori* collection of subgroups) with the largest generalized likelihood ratio test statistic and testing the hypothesis of no treatment effect for that subgroup, conditionally on the selection. Wager (2024b) sequentially split the data in order to sequentially construct and combine independent tests for treatment effect heterogeneity, with each subsequent test leveraging more data to learn the test statistic but using an independent fold to test. Fraser (1951) construct prediction sets of a desired coverage level using a procedure that iteratively and interactively cuts out regions of the sample space in a fashion that is reminiscent of chiseling. Indeed, a few of our intermediate lemmas recapitulate results in this work, though our inferential goal is quite different. The focus of Fraser (1951) is on characterizing the probability mass of the resulting region while ours is on leveraging the distribution of the samples within it.

A.4   Parametric and Bayesian approaches

Kovalchik et al. (2013) leverage a parametric "proportional interactions" model which assumes a structural relationship between the treatment response surface and the control response surface. Wan et al. (2024) construct confidence level sets for linear and generalized linear models by leveraging the known distribution of $\hat{\beta}$, the estimated regression coefficients. Though not precisely a parametric approach, Ma and Huang (2017) suppose that there are K subgroups that partition the data, and that units in the same subgroup have the same subject-specific intercept in a linear model with unobserved latent heterogeneity. They describe an optimization-based approach for recovering the subgroup structure and give conditions on the data-generating process (including K and the gap in subgroup effect between different partitions) under which the latent heterogeneity can be consistently recovered. Naturally, the validity of these approaches hinges on the validity of the model specification.

Bayesian approaches obviate selective inference concerns since the inferences are given conditionally on the data, on average over the prior. As with all Bayesian inference, validity depends upon having a well-specified prior and correctly interpreting the inference as an update to the prior in light of the data. A Bayesian model that has seen popular use in the assessment of treatment effect heterogeneity is the Bayesian additive regression tree, or BART, which also has a number of extensions specifically to the field of causal inference (Chipman et al., 2010; Hill, 2011; Hahn et al., 2020). BART is a fully Bayesian method that yields a posterior over the space of regression surfaces $\mathbb{E}[Y \mid X]$. Methods that consider Bayesian approaches more specifically for subgroup analysis include Berger et al. (2014), Schnell et al. (2016), and Dixon and Simon (1991). These methods either tend to be parametric, require the covariate factors to be dichotomous, and/or do not model interactions (or model a few, pre-specified interactions). An exception is Oganisian et al. (2020), which leverages an Enriched Dirichlet Process prior to model the space of "cost-effectiveness profiles."

This nonparametric prior naturally gives rise to clusters, which they interpret as subgroups. Some works investigate the intersection of Bayesian modeling and decision-theoretic aspects (Schnell et al., 2017). Finally, some works incorporate Bayesian ideas for testing and model selection, but ultimately evaluate their methodology according to frequentist criteria such as Type I error control (Sivaganesan et al., 2010). These methods typically do not yield rigorous frequentist inferential guarantees.

A.5 Conformal inference

Conformal inference (Vovk et al., 2005) provides distribution-free inferential guarantees by leveraging the exchangeability of the training and test samples, and providing guarantees that are unconditional over *both* training and test samples. The literature on conformal inference is now vast. We highlight three specific works that are most related to ours.

In causal inference, Lei and Candès (2021) construct conformalized prediction intervals for the individual treatment effect,

$$\mathbb{P}(Y_{n+1}(1) - Y_{n+1}(0) \in \hat{C}(X_{n+1})) \geq 1 - \alpha$$

where $\hat{C}$ is a function of samples $\{Y_i, T_i, X_i\}_{i=1}^n$ drawn from a randomized experiment, and $(Y_{n+1}(1), Y_{n+1}(0), X_{n+1})$ is an additional tuple drawn from the same distribution that the experiment is sampled from. We reiterate that the probability is taken over both $\{Y_i, T_i, X_i\}_{i=1}^n$ and $(Y_{n+1}(1), Y_{n+1}(0), X_{n+1})$. We also observe the inherent hardness of this problem, since the individual treatment effect is unidentifiable without assumptions on the joint distribution of the potential outcomes. Medarametla and Candès (2021) provide conformalized inferences for the conditional median $\text{Median}(Y \mid X = X_{n+1})$. This marks an interesting departure from previous works in conformal inference since the target of inference is a parameter of the distribution (which is never directly observed in the training sample) rather than the outcome itself. Jin and Candes (2023) specifically consider the task of selecting a proportion of units that satisfy a false discovery rate guarantee using ideas from conformal inference.

The guarantee that conformal inference provides is different from what we consider in this work, and in general they are difficult to compare. For one, a valid prediction interval does not directly imply anything about the expected welfare gain if one were to treat according to that prediction interval (for instance, treating all units whose prediction intervals exceed 0). Moreover, the coverage guarantees in conformal inference are unconditional over the training set, meaning that the realized population miscoverage rate can be far from the unconditional miscoverage rate (which is the nominal level by construction) unless certain algorithmic stability assumptions are satisfied (Liang and Barber, 2025).

Appendix B Simulation details and additional results

B.1 Type I error and exactness

The implementation of chiseling we consider here is the one discussed in Section 4.2. Let $f(x) = \arctan((X_1 + ... + X_5)/\sqrt{5})$ and in the case of RCTs always consider i.i.d. treatment indicators $W \sim \text{Bern}(0.5)$. We consider the following base simulation settings.

1. **Simulation Setting 1.** $X \in \mathbb{R}^{50}$ is a mean-centered correlated normal with covariance $\Sigma_{ij} = 0.2^{|i-j|}$. The machine learning algorithm used is ridge regression with leave-one-out cross-validation to select the penalty. We let $\epsilon \sim \text{Expo}(1) - 1$.

2. **Simulation Setting 2.** $X \in \mathbb{R}^{100}$ consists of independent Rademacher entries. The machine learning algorithm used is a multi-layer perceptron (neural network). We let $\epsilon \sim t_5$, the t-distribution with 5 degrees of freedom.

3. **Simulation Setting 3.** $X \in \mathbb{R}^{150}$ consists of independent $\text{Expo}(1) - 1$ entries. The machine learning algorithm used is random forests. We let $\epsilon \mid X \sim \mathcal{N}(0, 1 + f(X)^2)$ (i.e. the errors are heteroscedastic).

See Appendix B.2.2 for more details about the machine learning algorithms. Within each base setting, we consider three sub-settings. First, we consider the **binary** case where $Y \sim \text{Bern}(0.5)$ independently of X (and $\mu_{\text{cut}} = 0.5$). Next, we let $\mu_{\text{cut}} = 0$ and $Y'(1) = Y'(0) = f(X) + \epsilon$. We will consider both the **IPW transformation** and the **AIPW transformation**. The AIPW transformation is fit using 5-fold cross-fitting and linear regression on the first five coordinates, which is in particular misspecified.

In the first row of Figure B1, we empirically confirm the theoretical exactness of chiseling when Y is binary (using the exact test for binary outcomes designed in Section 3.6). Next, we run the asymptotic variant of chiseling while only enforcing a minimum sample size of $n_{\text{min}} = 30$ in order to assess the robustness of our asymptotic approximation. We find that even with such a small minimum sample size, our asymptotic approximation is of high quality and the Type I error is essentially exact at level α (second and third rows of Figure B1). Note, importantly, that the asymptotic regime is determined not by the sample size n but by the minimum sample size n_{min}. Because in these experiments we are not scaling n_{min} with n, we are not approaching an asymptotic regime as n grows. Nonetheless, the Type I error is robust, and the asymptotics can only improve as n_{min} is increased (for instance, requiring it to be 5% of n).

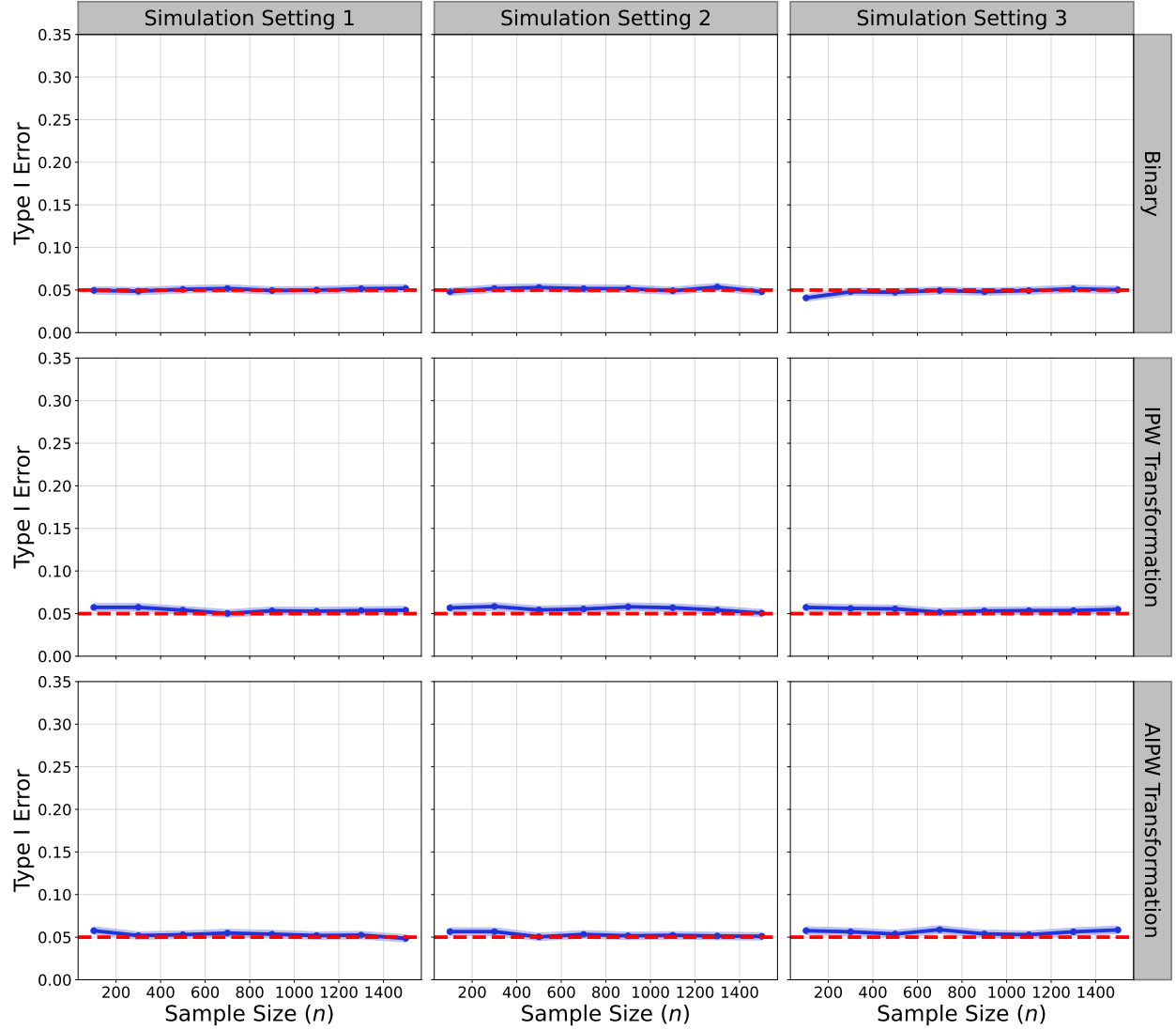Figure B1: Type I error of chiseling under the global null across different data generating processes and machine learning methods (target level is $\alpha = 0.05$, given by the dashed red line). Error band shown reflects $\pm$ twice the standard error. A minimum sample size of $n_{\min} = 30$, not scaling with n, is enforced in order to showcase the robustness of chiseling to small $n_{\min}$.

B.2 Simulation details

B.2.1 p-simultaneous data splitting

The dataset is split into a training set of size $\lfloor pn \rfloor$, with the rest set aside for testing. Using the training dataset, a machine learning model is used to estimate $\hat{\mu}(\cdot)$ and $\mathcal{R}_1 := \{x : \hat{\mu}(x) > \mu_{\text{cut}}\}$. We then define $\mathcal{R}_2, ..., \mathcal{R}_{10}$ by thresholding $\hat{\mu}(\cdot)$ at different cutoffs, where the cutoffs are chosen so that $\mathcal{R}_{10}$ is the smallest region with at least $n_{\min}$ samples and the number of samples in $\mathcal{R}_j \setminus \mathcal{R}_{j-1}$ for $j = 2, ..., 10$ is as even as possible. Though we define these regions using the test data, we will treat them as fixed. This approximation is reasonable since they are only defined using X.

When the outcomes are continuous, we let $\hat{\theta}_j$ denote estimates of $\theta_j := \mu(\mathcal{R}_j)$ obtained by taking the simple mean of points in the region. We then calculate the standard error $\hat{\sigma}_j$ for each $\hat{\theta}_j$ by bootstrapping the test sample. We form the statistics

$$T := \max_j \left\{ \frac{\hat{\theta}_j - \theta_j}{\hat{\sigma}_j} \right\}$$

and in particular B bootstrapped versions of the above $T^{(1)}, ..., T^{(B)}$. The lower simultaneous confidence bounds are then given by $\hat{\theta}_j - \hat{\sigma}_j q$ for $j = 1, ..., 10$ where q is the $1 - \alpha$ quantile of $\{T^{(1)}, ..., T^{(B)}\}$. We report $\mathcal{R}_j$ for the smallest j such that $\hat{\theta}_j - \hat{\sigma}_j q > \mu_{\text{cut}}$. The validity of this construction follows from standard theory regarding the validity of the bootstrap for Z-estimators (e.g. Chapter 10 of Kosorok (2008)) once we recognize that $(\hat{\theta}_1, ..., \hat{\theta}_{10})$ is the solution to a set of Z-estimating equations.

When the outcomes are binary, we are not aware of any obvious method that provides simultaneous testing for the means of nested regions that is valid in finite samples. We find that the asymptotic approximation can fail in small samples, so instead we consider the following approach. We form the natural statistic

$$S_j := \sqrt{n(\mathcal{R}_j)} \cdot \frac{\hat{\theta}_j - \mu_{\text{cut}}}{\sqrt{\hat{\theta}_j (1 - \hat{\theta}_j)}}$$

where $n(\mathcal{R}_j)$ is the number of test samples in $\mathcal{R}_j$. If $\hat{\theta}_j = 1$ we let $S_j = \infty$, and if the right-hand side is otherwise undefined we let it equal 0. This is the Z-statistic for testing $H_j : \theta_j \leq \mu_{\text{cut}}$ where we have imputed some edge cases. We will reject H_j for all $S_j > c$. How should we set c? We condition on X (which in particular fixes the regions) and desire

$$\mathbb{P}_{\mathcal{P}}(\exists j : S_j > c \text{ and } \mathcal{P} \in H_j \mid X) \leq \alpha$$

for all distributions $\mathcal{P}$. Let $c(\mathcal{P})$ be the smallest c satisfying the above for distribution $\mathcal{P}$. Ideally, we would take $c := \sup_{\mathcal{P}} c(\mathcal{P})$, which would ensure the finite-sample validity of our test. As an *anti-conservative* approximation, we let $\tilde{c} := c(\mathcal{P}_0)$ where $\mathcal{P}_0$ is the distribution where $Y_1, ..., Y_n$ are i.i.d. $\text{Bern}(\mu_{\text{cut}})$ and independent of X. Clearly, $\tilde{c} \leq c$, and so this approximation rejects strictly more often than the finite-sample valid test. Furthermore, we can easily calculate $\tilde{c}$ by simulations. This lets us *upper bound* the expected utility and power of the finite-sample variant of p-simultaneous data splitting, and this is what we show

throughout our simulations. Though it is an upper bound, we actually expect it to be quite close to the power of the finite-sample valid variant, as $\mathcal{P}_0$ is specifically chosen with the heuristic that it is at least close to maximizing $c(\mathcal{P})$ (it is at the boundary of the global null). To the extent that it fails to maximize $c(\mathcal{P})$, it creates an inflation in expected utility and power. Lastly, we note that in this variant, it can be quite important to randomize the quantile since the test statistics are highly discrete. Thus, we in fact let $\tilde{c}$ be the randomized quantile such that $\mathbb{P}_{\mathcal{P}_0}(\exists j : S_j > \tilde{c} \mid X) = \alpha$ exactly.

B.2.2 Machine learning methods

We use off-the-shelf implementations from scikit-learn of ℓ^2-penalized logistic regression, lasso with 5-fold cross-validation, ridge regression with leave-one-out cross-validation, multi-layer perception regression (i.e. neural network), and random forest regression and classification. The only modifications we make are that for the linear model-based methods we standardize X (and Y, where appropriate) prior to fitting, and that for ridge regression we increase the grid of penalty parameters to 20 evenly spaced values between 10^{-5} and 10^5 in log-space. Unless stated otherwise, when the IPW transformation is used to convert a causal problem into a regression problem, the resulting pseudo-outcomes are directly regressed onto the covariates. For computational reasons, when chiseling we generally do not refit the machine learning model each time a new data point is revealed, but only in batches of 1%, 5% or 10% depending on the computational cost of the simulation setting.

B.2.3 Data generating processes

Throughout, when the simulation is an RCT we let $Y'(1) = \mu(X) + Y'(0)$ where we will specify $\mu(\cdot)$ and $Y'(0)$. The treatment indicator will always be i.i.d. $W \sim \text{Bern}(0.5)$, and we always work with the regression formulation with IPW outcomes, i.e. $Y = 2WY'(1) - 2(1 - W)Y'(0)$. Machine learning methods are directly regressed on the IPW outcomes.

DGPs for Section 3.4. For the linear case, we let $X \sim \mathcal{N}(0, I_d)$, $\mu(X) = X^T\beta$, and $Y(0) \sim \mathcal{N}(0, 1)$. We let $d = 20$ and $\beta \propto (1, ..., 1)$ with $\|\beta\| = 0.5$. For the non-linear case, we let X and $\mu(X)$ follow the data generating process of Kang and Schafer (2007), to which we defer for additional details. The noise term in Kang and Schafer (2007) is $\mathcal{N}(0, 1)$, but we inflate the noise so that $Y(0) \sim \mathcal{N}(0, 80^2)$. Furthermore, we set the intercept so that $\mathbb{E}[Y(1)] = 0$. Note that the Kang and Schafer (2007) example also has non-constant propensities, which we ignore since we are imitating an RCT.

DGP for Section 5.2. We let $X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{d \times d}$ and $\Sigma_{ij} = 0.2^{|i-j|}$. We let $\mathbb{P}(Y = 1 \mid X) = \mu(X) = \text{logit}(\tau + X^T\beta)$. We let $d = 100$, $\beta \propto (1, 1, 1, 1, 1, 0, ..., 0)$. Letting $\|\beta\| = \theta$, we consider $(\theta, \tau) \in \{(2, 2.2), (1.3, 0.243), (1.5, -1.936)\}$ so that the probability masses of the optimal regions are 0.5, 0.1, and 0.01 respectively.

DGP for Section 5.3. We let $X \sim \mathcal{N}(0, \Sigma)$ where $\Sigma \in \mathbb{R}^{d \times d}$ and $\Sigma_{ij} = 0.2^{|i-j|}$. We let $\mu(X) = \tau + X^T\beta$ and $Y(0) \sim \text{Expo}(1) - 1$. We let $d = 100$, $\beta \propto (1, 1, 1, 1, 1, 0, ..., 0)$. Letting

$\|\beta\| = \theta$, we consider $(\theta, \tau) \in \{(0.45, 0), (0.4, -0.601), (0.8, -2.201)\}$ so that the probability masses of the optimal regions are 0.5, 0.1, and 0.01 respectively.

DGP for Section 5.4. We let $\mu_{\text{cut}} = 0$ and again consider the DGP from Kang and Schafer (2007). We now let $Y(0) \sim \mathcal{N}(0, 125^2)$ and vary $\tau \in \{24.45, 0, -24.35\}$ so that the probability masses of the optimal subgroups are $0.75, 0.5,$ and 0.25 respectively. Here, the CATE is nonlinear in X but we use a linear model to learn the subgroups.

DGP for Section 5.5. First we draw $Z \sim \text{Bern}(q)$. Then we draw $L \in \mathbb{R}^5$ from the distribution of $A \mid A_1 + ... + A_5 = 0$ where $A \sim \mathcal{N}(0, I_5)$. Then we define $X_j = \tau Z + \tau_0 (1 - Z) + L_j$ for $j = 1, ..., 5$ and let $\mu(X) = (X_1 + ... + X_5)/5 = \tau Z + \tau_0(1 - Z)$. In the first simulation setting, we let $\tau_0 = 0$. In the second, we let $\tau_0 = -0.1\tau$. We let $\tau = 0.1/q$. The sample size is $n = 1000$.

B.3 Additional results
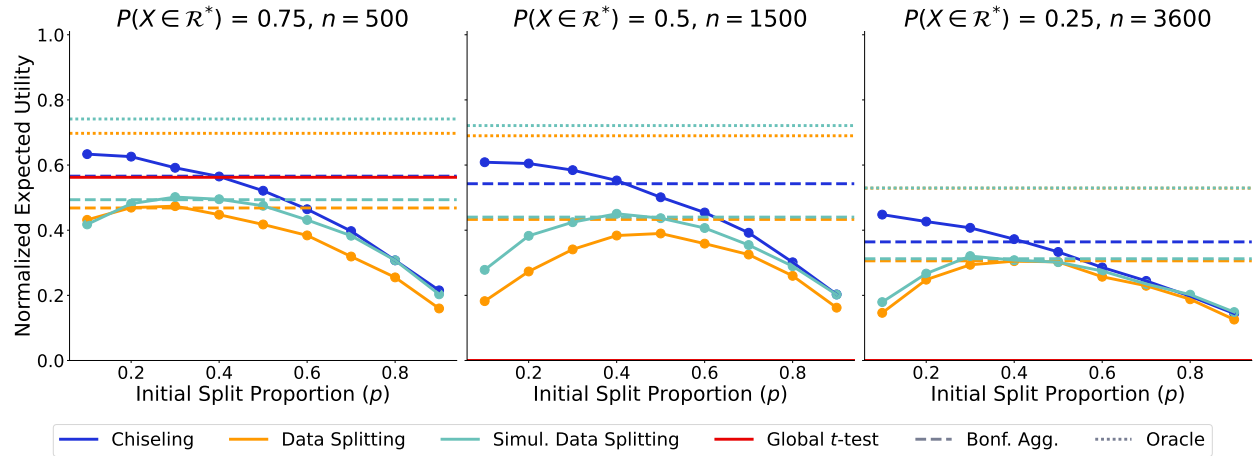

Figure B2: Results for Section 5.4. Normalized expected utility in a simulated RCT where the CATE is non-linear. Solid horizontal lines correspond to using Bonferroni to aggregate across initial split proportions $\in \{0.2, 0.5, 0.8\}$ for the method with the corresponding color. The machine learning method used is ridge regression with leave-one-out cross-validation, which is misspecified.

Table 1: **Top 10 features determining chiseled subgroup membership in GSS dataset.** "Rate" refers to the proportion of units who would be included in the subgroup based on that feature alone, and features are ranked by rate. Some features are redundant because they represent different codings of the same survey item.

| Rank	Feature ID	Feature Description	Inclusion Criteria	Rate
1	partyid	Political party affiliation	Neither, no response, independent, close to Republican, not very strong Republican, strong Republican, or other party	53.2%
2	year	GSS year for respondent	1998 or earlier	53.5%
3	polviews	Political ideology	Moderate, slightly conservative, conservative, or extremely conservative	68.3%
4	income	Total family income	≥$15,000	68.9%
5	income_num	Total family income	≥$15,000	68.9%
6	racdif3	Differences due to education	Did not answer "yes" to "On the average African-Americans have worse jobs, income, and housing than white people. Do you think these differences are because most African-Americans don't have the chance for education that it takes to rise out of poverty?"	72.6%
7	marital	Marital status	Married, widowed, or divorced	73.7%
8	polviews_num	Political ideology	Moderate, slightly conservative, conservative, or extremely conservative	76.1%
9	racdif1	Differences due to discrimination	Did not answer "yes" to "On the average African-Americans have worse jobs, income, and housing than white people. Do you think these differences are mainly due to discrimination?"	78.1%
10	race	Race of respondent	White	79.1%
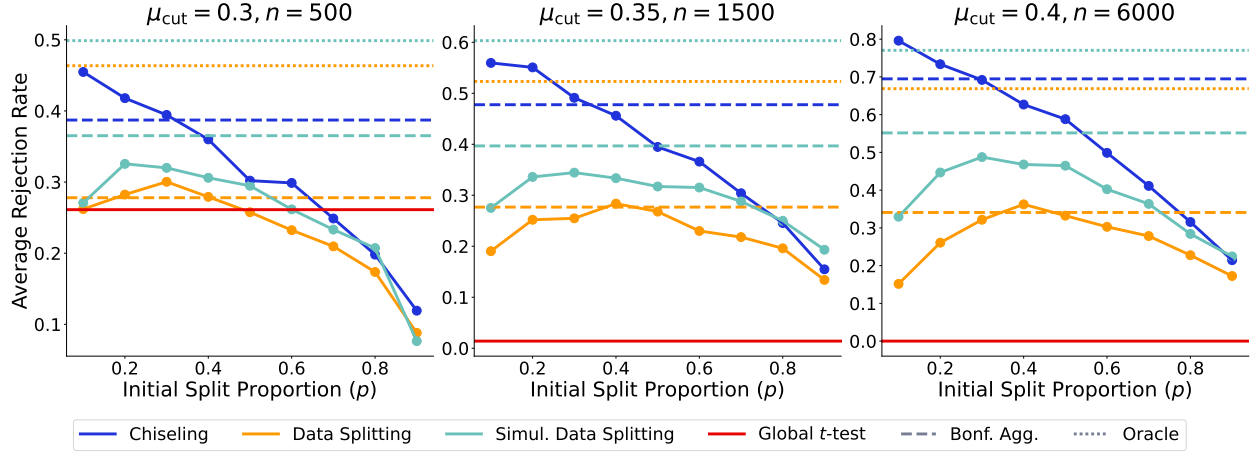
Figure B3: Results for Section 6. The analogue of Figure 6 but for the average rejection rate across repeated train/validation splits.
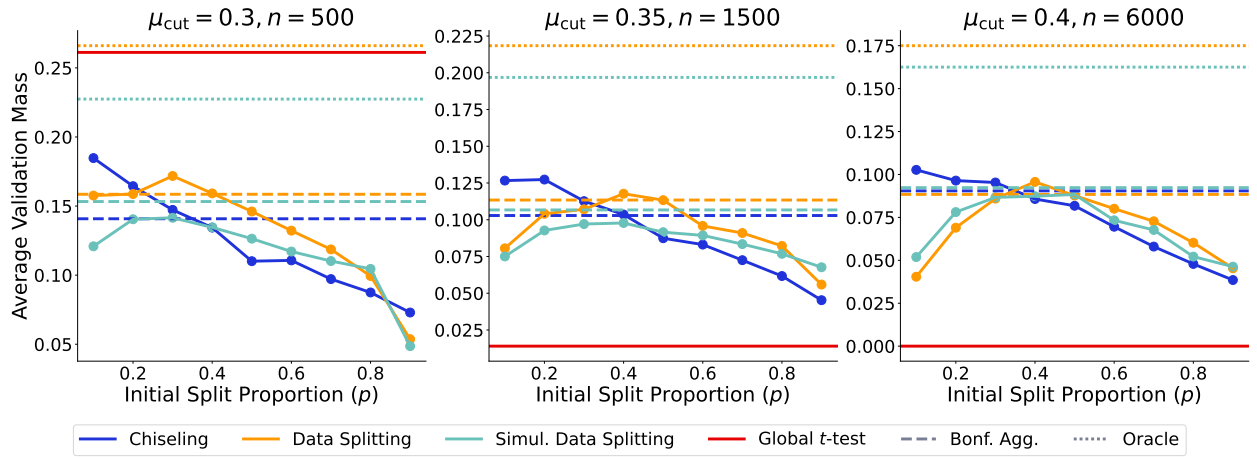

Figure B4: Results for Section 6. The analogue of Figure 6 but for the average probability mass of the rejected region (estimated using the validation set) across repeated train/validation splits.
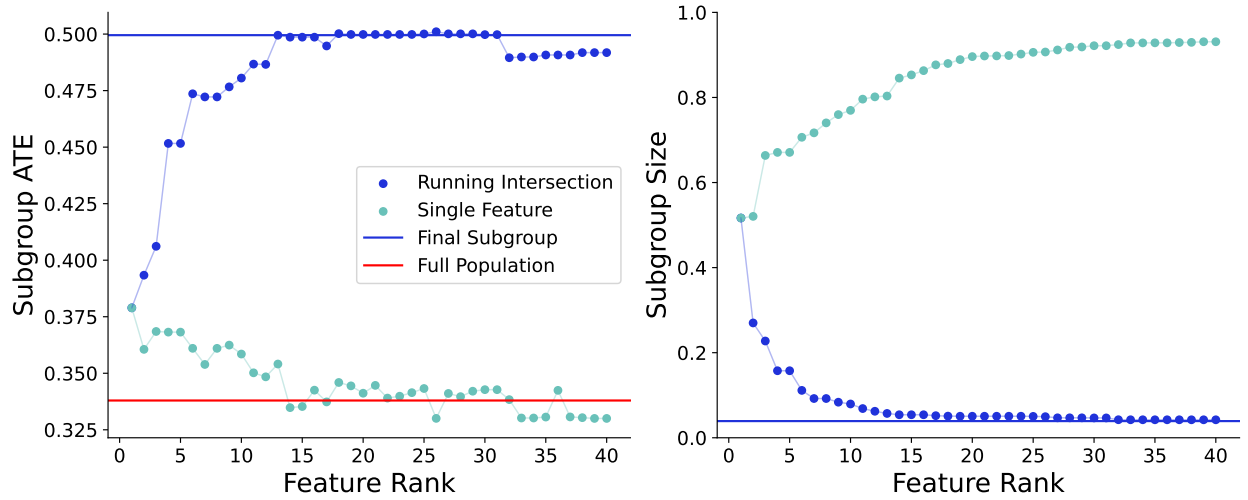
Figure B5: The rejected region has the form $[l_1, u_1] \times ... \times [l_d, u_d]$, and feature j's importance is given by the fraction of samples falling in $[l_j, u_j]$. **Left panel:** the estimated subgroup ATE within subgroups defined by single feature inclusion criteria, as well as the estimated subgroup ATE within the subgroup defined by the running intersection of inclusion criteria (both in decreasing order of feature importance) for the first 40 features. Intersecting all $d = 152$ inclusion criteria yields the subgroup discovered by chiseling. We also show the full population ATE for comparison. **Right panel:** the same as the left panel but where subgroup ATE is replaced with subgroup probability mass (i.e. the proportion of units belonging to the subgroup).

B.4 Multiple testing chiseling versus single test chiseling
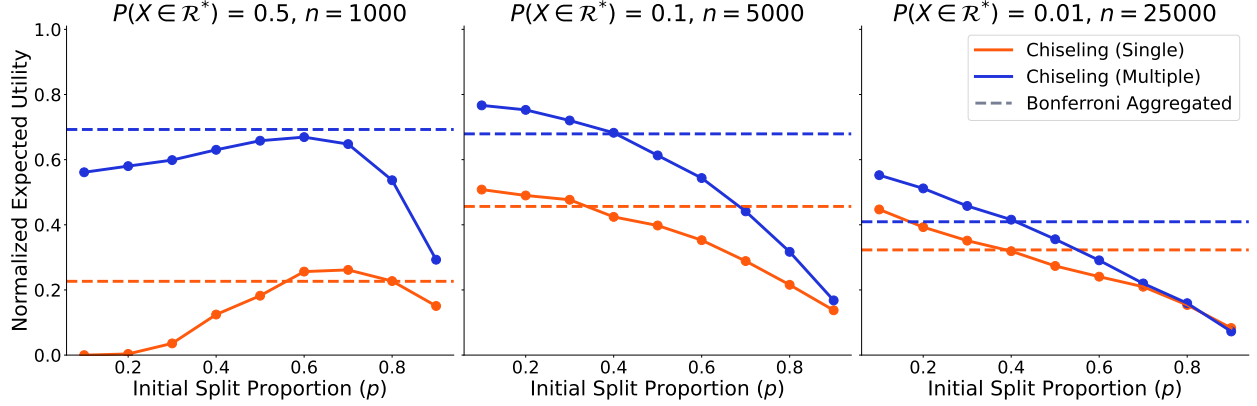
Figure B6: A comparison of chiseling with multiple testing (Section 4.2) versus chiseling with a single test (Section 3.4) in the simulation setting of Section 5.2.
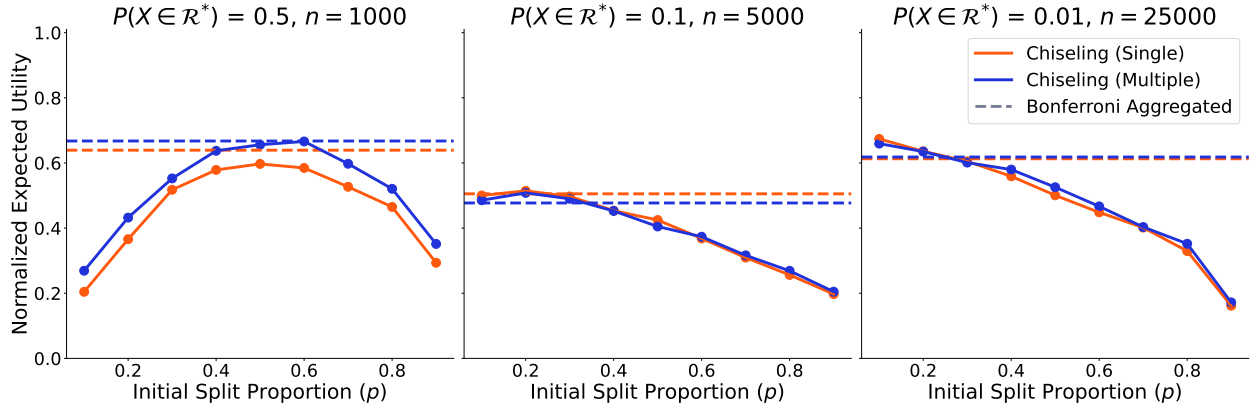
Figure B7: A comparison of chiseling with multiple testing (Section 4.2) versus chiseling with a single test (Section 3.4) in the simulation setting of Section 5.3.
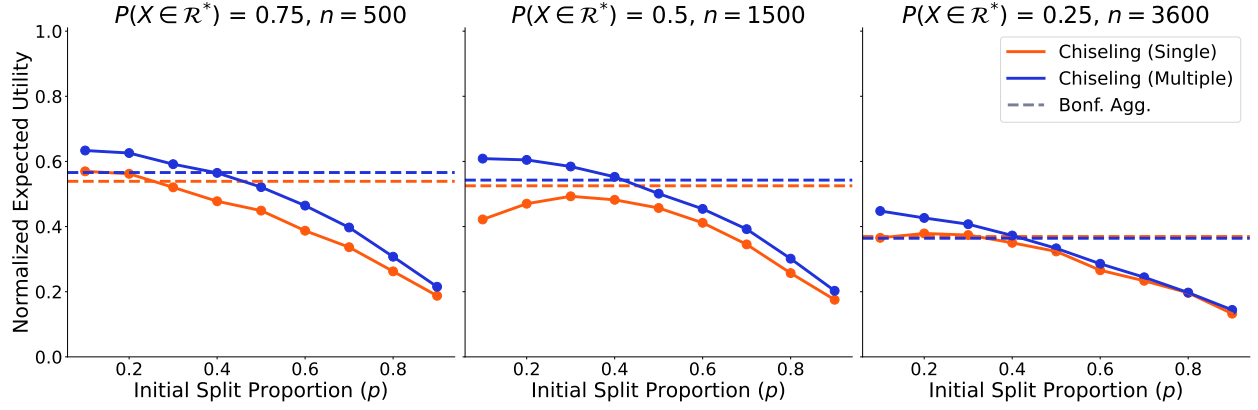
Figure B8: A comparison of chiseling with multiple testing (Section 4.2) versus chiseling with a single test (Section 3.4) in the simulation setting of Section 5.4.

B.5 Power
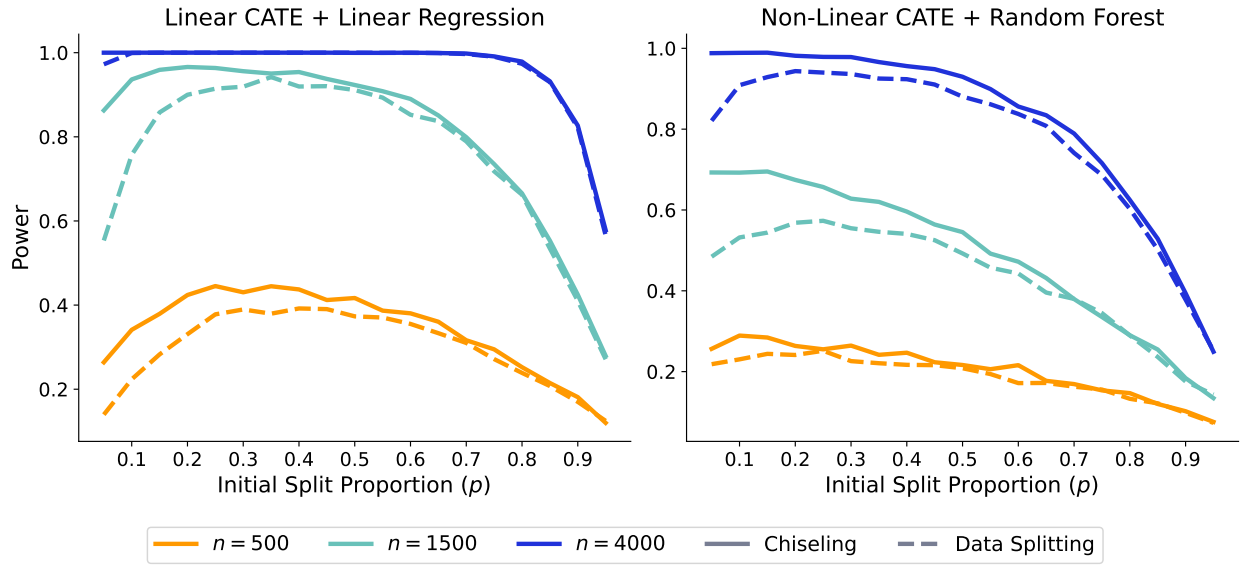
Figure B9: Results for Section 3.4. The analogue of Figure 2 for power.
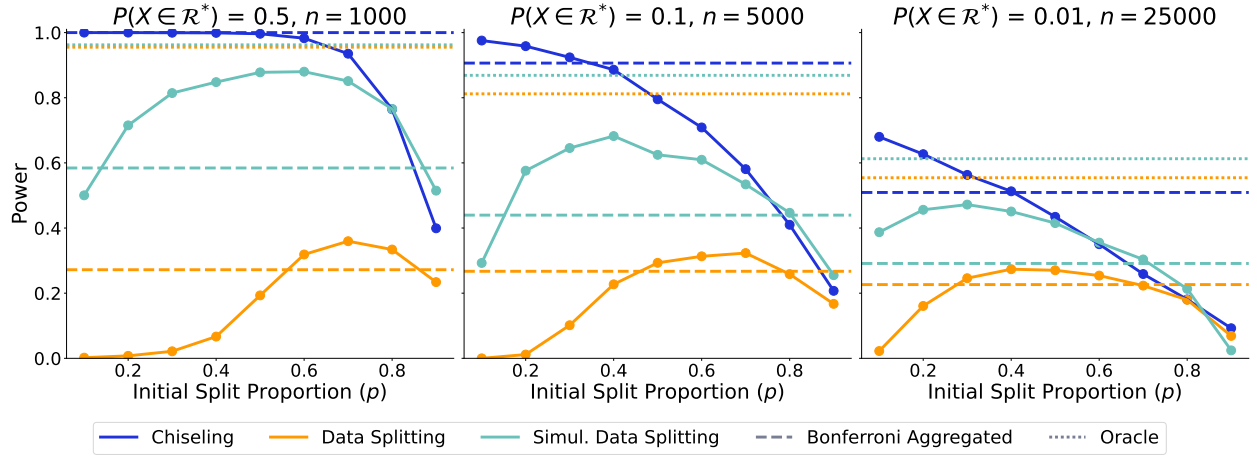
Figure B10: Results for Section 5.2. The analogue of Figure 3 for power.
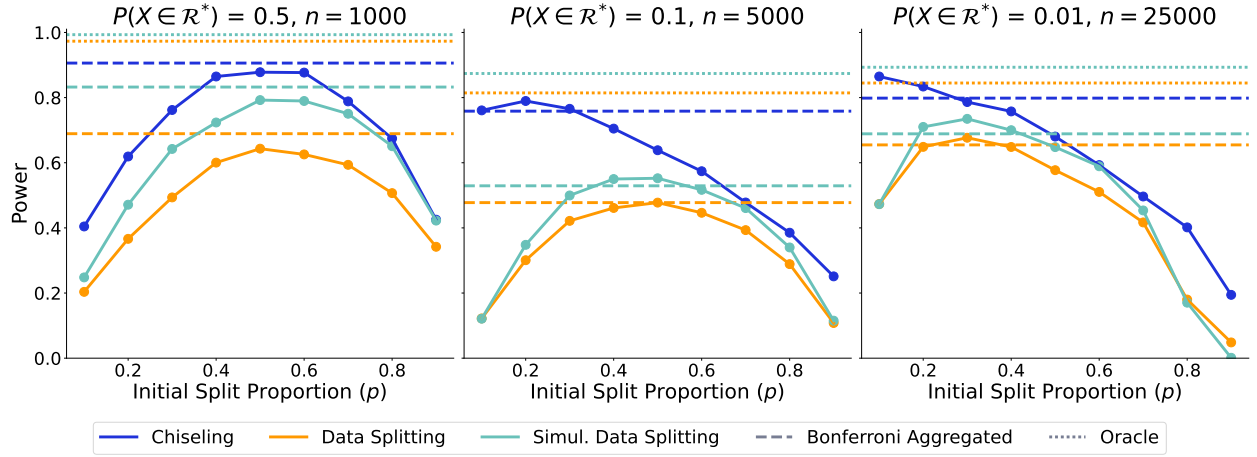
Figure B11: Results for Section 5.3. The analogue of Figure 4 for power.
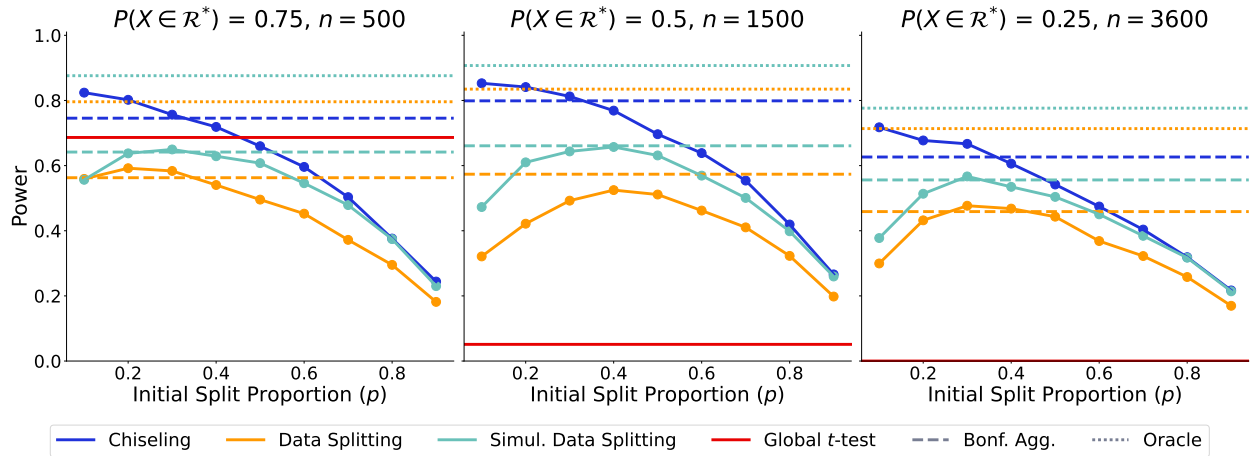
Figure B12: Results for Section 5.4. The analogue of Figure B2 for power.
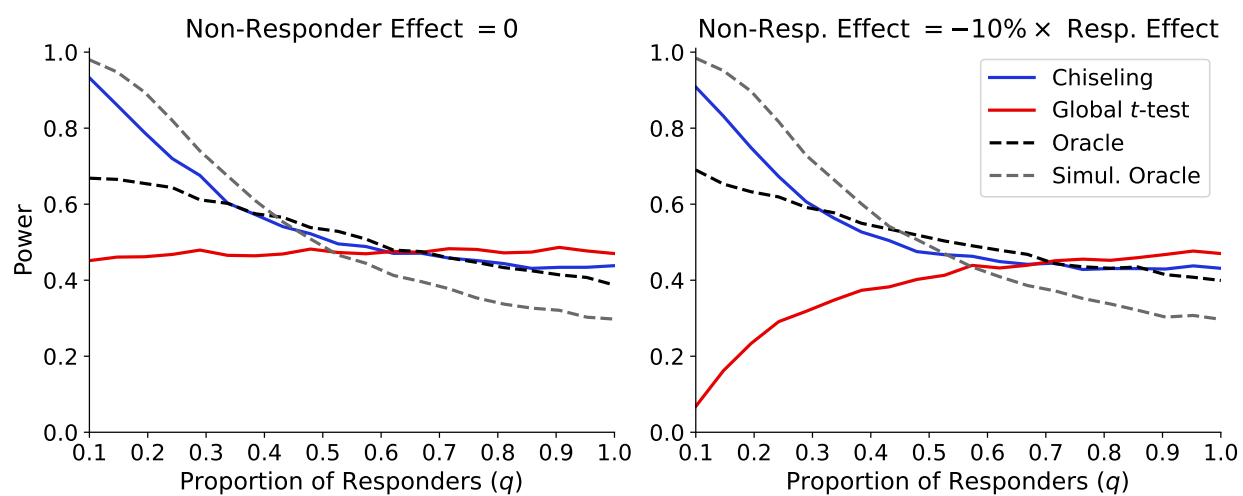
Figure B13: Results for Section 5.5. The analogue of Figure 5 for power.

Appendix C   Extended discussion

C.1   Relationship between utility and policy value

In the policy learning literature, it is common to let $\mu_{\text{cut}} = 0$ (to protect against discovering non-efficacious treatment policies) and to desire policies with high *policy value*. This is equivalent to Equation (4) since

$$
\begin{aligned}
\mathbb{E}[Y \mathbb{1}\{X \in \mathcal{R}\}] &= \mathbb{E}[(Y'(1) - Y'(0))\mathbb{1}\{X \in \mathcal{R}\}] \\
&= \mathbb{E}[Y'(1)\mathbb{1}\{X \in \mathcal{R}\} + Y'(0)\mathbb{1}\{X \notin \mathcal{R}\}] - \mathbb{E}[Y'(0)] \\
&= \mathbb{E}[Y'(\pi(X))] - \mathbb{E}[Y'(0)]
\end{aligned}
$$

where $\pi(X) = \mathbb{1}\{X \in \mathcal{R}\}$ is the policy that treats everyone in $\mathcal{R}$. Thus, Equation (4) is equivalent to the policy value $\mathbb{E}[Y'(\pi(X))]$ up to the constant factor $\mathbb{E}[Y'(0)]$. Thus, one maximizer of the policy value is the policy that treats units whose CATEs are positive, i.e., the units in $\mathcal{R}^*$.

C.2   Querying regions

For $t \geq 1$, let $f_t(\cdot)$ and c_t denote what is chosen by the tth application of Algorithm 2 used to generate $(\mathcal{R}_t)_{t=0}^{\infty}$. Let $\gamma_t := c_t \wedge \min_{i:X_i \in \mathcal{R}_{t-1}} f_t(X_i)$. For any $x \in \mathcal{X}$, one may check whether $x \in \mathcal{R}_t$ by simply checking whether $f_s(x) > \gamma_s$ for all $0 < s \leq t$. If the scoring functions $f_t(\cdot)$ are computationally feasible to evaluate, then it is feasible to check whether any x belongs to $\mathcal{R}_t$ or not. In the context of causal inference, this makes the treatment rule "treat a person with covariates x if $x \in \mathcal{R}_t$" one that is computationally feasible to use in practice.

C.3   Total order

In Algorithm 2, we have chosen to parameterize the shrinking of regions by the upper level sets of a real-valued function as this is a simple, intuitive, and useful choice. However, it is not the most general parameterization. In general, $f(\cdot)$ may take values in any set that is given a total order—a binary relation that is reflexive, transitive, antisymmetric, and connected. For example, the analyst may choose $f : \mathcal{X} \to \mathbb{R}^2$ and give $\mathbb{R}^2$ the lexicographic order. Then in Algorithm 2 we may replace minima with lexicographic minima and comparisons such as $>$ with lexicographic comparisons; Theorem 3.2 and Corollary 3.3 will still hold.

C.4   Chiseling generalizes data splitting

Suppose without loss of generality that the covariates have the form $(X, U) \in \mathcal{X} \times [0, 1]$ where U is an independent standard uniform random variable (appending an extra feature if necessary). Then for any region of the form $\mathcal{R} = \mathcal{R}' \times I$ for $\mathcal{R}' \subseteq \mathcal{X}$ and $I \subseteq [0, 1]$, we have $\mathbb{E}[Y \mid (X, U) \in \mathcal{R}] = \mathbb{E}[Y \mid X \in \mathcal{R}']$. Thus, guarantees on the former translate to guarantees on the latter. Now as long as either

$$
f(x, u) = f'(x) \text{ for some } f' \quad \text{or} \quad f(x, u) = u
$$

in each application of Algorithm 2, we will have that all $\mathcal{R}_t$ are of the form $\mathcal{R}'_t \times (u, 1]$ for some $\mathcal{R}'_t \subseteq \mathcal{X}$. The former choice is used to shrink the region within the covariate space of

interest $\mathcal{X}$, while the latter is used to reveal a point uniformly at random from the current region. Splitting off k random data points is thus equivalent to chiseling with $f(x, u) = u$ successively k times. Moreover, this choice may be used to reveal points randomly from the current region at *any* point during the procedure.

We note that here since $\mathcal{R}'_t$ are the regions of interest, though the ambient space is $\mathcal{X} \times [0, 1]$ we should redefine utility as $\mathcal{U}(\mathcal{R}') = \mathbb{E}[(Y - \mu_{\text{cut}})\mathbb{1}\{X \in \mathcal{R}'\}]$ for $\mathcal{R}' \subseteq \mathcal{X}$ (i.e. we do not marginalize over U).

C.5 Tiebreaking

If multiple data points in the region have the same $f(\cdot)$ value, then they will all be revealed by Algorithm 2. If the analyst wishes for each application of Algorithm 2 to reveal no more than one data point, then this can be accomplished by randomly breaking ties in $f(\cdot)$. In fact, this can be formalized by adding an additional auxiliary dimension of independent standard uniforms (as in Appendix C.4) and letting $\tilde{f}(x, u) = (f(x), u)$, now using lexicographic comparisons to order points (Appendix C.3). Then this will reveal points in order of $f(x)$ but break ties using u. However, it results in a region that depends in a more complex way on the auxiliary random feature, and hence in the context of causal inference corresponds to a treatment rule that stochastically treats a fraction of units at the boundary of the region. Whether this is useful in practice or not is case-dependent. However, we will use this tiebreaking construction in the proof of Lemma E.11.

C.6 Chiseling to the boundary

Algorithm 3 is the concrete algorithm that produces the regions used to compare chiseling to data splitting in Section 3.4. It also corresponds to the first step of the strategy described in Section 4.2. Note that Line 5 corresponds to a stopping time, so that if we let ν count the number of times Algorithm 2 is called, then Corollary 3.3 indeed applies to the output $\mathcal{R}_\nu$.

Algorithm 3 Chiseling to the boundary using machine learning

      Input: dataset $\mathcal{D} = (X_i, Y_i)_{i=1}^n$, region $\mathcal{R}$, split proportion $p \in (0, 1)$, ML method

1: $\mathcal{R}_\nu \leftarrow \mathcal{R}, \mathcal{M}_\nu \leftarrow$ random subset of $[n]$ of size $\lceil (1 - p)n \rceil$
2: **while** $|\mathcal{M}_\nu| > 0$ **do**
3:     Refit $\hat{\mu}(\cdot)$ to $(X_i, Y_i)_{i \notin \mathcal{M}_\nu}$ using ML method to predict $\mu(x) = \mathbb{E}[Y \mid X = x]$
4:     Obtain $\tilde{\mathcal{R}} \subseteq \mathcal{R}_\nu$ by chiseling using Algorithm 2 with $f(\cdot) = \hat{\mu}(\cdot)$ and $c = \mu_{\text{cut}}$
5:     **if** $|\{i \in \mathcal{M}_\nu : X_i \in \mathcal{R}_\nu \setminus \tilde{\mathcal{R}}\}| = 0$ **then** quit loop after **update**
6:     **update** $(\mathcal{R}_\nu, \mathcal{M}_\nu) \leftarrow (\tilde{\mathcal{R}}, \mathcal{M}_\nu \setminus \{i \in \mathcal{M}_\nu : X_i \in \mathcal{R}_\nu \setminus \tilde{\mathcal{R}}\})$
7: **end while**
8: **Return:** $\mathcal{R}_\nu, \mathcal{M}_\nu$

C.7 Derivation of conditioning–truncation equivalence

Recall the notation from Section 3.6. Then for $s < t$,

$$n_s \bar{\mu}_s - n_t \bar{\mu}_t = \sum_{i:X_i \in \mathcal{R}_s} Y_i - \sum_{i:X_i \in \mathcal{R}_t} Y_i = \sum_{i:X_i \in \mathcal{R}_s \setminus \mathcal{R}_t} Y_i$$

which, by rearranging, implies that

$$\bar{\mu}_t = \frac{1}{n_t}\left(n_s \bar{\mu}_s - \sum_{i:X_i \in \mathcal{R}_s \setminus \mathcal{R}_t} Y_i\right).$$

Then,

$$\bar{\mu}_s \le C_s \iff \bar{\mu}_t \le \frac{1}{n_t}\left(n_s C_s - \sum_{i:X_i \in \mathcal{R}_s \setminus \mathcal{R}_t} Y_i\right).$$

Therefore, $\max_{s<t} \phi_s = 0$ if and only if

$$\bar{\mu}_s \le C_s \text{ for all } s < t \iff \bar{\mu}_t \le \min_{s<t}\left\{\frac{1}{n_t}\left(n_s C_s - \sum_{i:X_i \in \mathcal{R}_s \setminus \mathcal{R}_t} Y_i\right)\right\} \iff \bar{\mu}_t \le M_t.$$

C.8   Truncated binomial quantiles

Recall in Definition 3.8 that we defined $Q(q; \tilde{n}, \tilde{\mu}, \tilde{M})$ to be the qth quantile of $\text{Binom}(\tilde{n}, \tilde{\mu})$ truncated to be no greater than $\tilde{M}$. However, due to the discreteness of the binomial distribution, an exact qth quantile may not exist. In this case, $Q(q; \tilde{n}, \tilde{\mu}, \tilde{M})$ returns an exogenously randomized quantile. Concretely, let $Z \sim \text{Binom}(\tilde{n}, \tilde{\mu})$. Then let

$$z^*_{\text{upper}} := \inf_{z \in \mathbb{N}}\left\{\mathbb{P}(Z \le z \mid Z \le \tilde{M}) > q\right\},$$
$$z^*_{\text{lower}} := \sup_{z \in \mathbb{N}}\left\{\mathbb{P}(Z \le z \mid Z \le \tilde{M}) \le q\right\}.$$

Then define

$$Q(q; \tilde{n}, \tilde{\mu}, \tilde{M}) := \begin{cases} z^*_{\text{upper}} & \text{with probability } \frac{q - \mathbb{P}(Z \le z^*_{\text{lower}} \mid Z \le \tilde{M})}{\mathbb{P}(Z \le z^*_{\text{upper}} \mid Z \le \tilde{M}) - \mathbb{P}(Z \le z^*_{\text{lower}} \mid Z \le \tilde{M})}, \\ z^*_{\text{lower}} & \text{otherwise.} \end{cases}$$

In other words, $Q(q; \tilde{n}, \tilde{\mu}, \tilde{M})$ is a random variable that we always consider to be completely independent of all other random variables. Under this definition,

$$\mathbb{P}\left(Z \le Q(q; \tilde{n}, \tilde{\mu}, \tilde{M}) \mid Z \le \tilde{M}\right) = q.$$

In Algorithm 1, we imagine that each time this function is queried, it produces a random output, but that once the output is observed, its value is fixed. Though verbose, all of these quantities can be readily and efficiently calculated given access to the binomial CDF. Lastly, note that if $\mathbb{P}(Z \le z^*_{\text{lower}} \mid Z \le \tilde{M}) = q$, then $Q(q; \tilde{n}, \tilde{\mu}, \tilde{M}) = z^*_{\text{lower}}$ with probability 1. So deterministic versions of our methods can be implemented by choosing q so that this equality can be achieved. In practice, if α_t is the desired error level, we recommend rounding α_t to the nearest value such that a deterministic quantile will be reported.

C.9 Derivation of critical values

Let us consider the case of binary Y. As per Lemma 3.7 and Appendix C.7, we need to construct C_t so that

$$\mathbb{P}(\bar{\mu}_t > C_t \mid \mathcal{F}_t, \{\mu(\mathcal{R}_t) \leq 0, \bar{\mu}_t \leq M_t\}) \leq \alpha_t \quad \text{a.s.}$$

Conditional on $\mathcal{F}_t$ only, the binary nature of Y means that $n_t \bar{\mu}_t \sim \text{Binomial}(n_t, \mu(\mathcal{R}_t))$. When $\mathcal{R}_t$ is null (i.e. $\mu(\mathcal{R}_t) \leq \mu_{\text{cut}}$), then $\text{Binomial}(n_t, \mu(\mathcal{R}_t))$ is stochastically dominated by $\text{Binomial}(n_t, \mu_{\text{cut}})$. But we must additionally condition on $\bar{\mu}_t \leq M_t$, which is equivalent to $n_t \bar{\mu}_t \leq n_t M_t$. In general stochastic domination is not necessarily preserved under right truncations, but in the case of binomials it is preserved (Lemma F.6). This motivates letting $n_t C_t$ be the $1 - \alpha_t$ quantile of an appropriately truncated binomial distribution, which yields the critical value in Definition 3.8.

Note that since $\sum_{i:X_i \in \mathcal{R}_s \setminus \mathcal{R}_t} Y_i$ and n_s are $\mathcal{F}_t$-measurable for all $s < t$, then an inductive argument implies that M_t is $\sigma(\mathcal{F}_t, (U_s)_{s<t})$-measurable for all t, where U_t is some auxiliary random variable, independent of everything, used to generate the tth randomized quantiles outputted by $Q(q; \tilde{n}, \tilde{\mu}, \tilde{M})$ which are used to define C_t (see Appendix C.8 for more details on the randomized quantiles). This measurability implies that the truncation level M_t can be calculated without peeking inside $\mathcal{R}_t$. This important property allows us to characterize the distribution of the sample means under the null without requiring any information from inside the region, thus preventing any further distortion of the distribution of the masked sample.

The story for general Y is similar, except that M_t is no longer $\mathcal{F}_t$-measurable due to the fact that $\hat{V}_t$ depends on data from inside the region. If we replaced $\hat{V}_t$ with $\sigma^2(\mathcal{R}_t)$ where $\sigma^2(\mathcal{R}) := \text{Var}(Y \mid X \in \mathcal{R})$, then M_t would in fact be $\mathcal{F}_t$-measurable, and this makes it easier to characterize the null distribution of each test statistic conditional on having not previously rejected. In Appendix D.4, we prove validity for an oracle version of our procedure which uses $\sigma^2(\mathcal{R}_t)$ in place of $\hat{V}_t$, and then prove that these two procedures produce the same output with probability converging to 1 asymptotically.

C.10 Clipped critical value and non-exactness

As mentioned in Section 3.6, in Definition 3.10 we clip C_t to be no smaller than 0. This is to address an asymptotic subtlety: even if $\sup_z |\mathbb{P}(Z_n \leq z) - \Phi(z)| \to 0$, we may not have $\sup_{z \leq u_n} |\mathbb{P}(Z_n \leq z)/\mathbb{P}(Z \leq u_n) - \Phi(z)/\Phi(u_n)| \to 0$ if u_n diverges to $-\infty$ too fast. Here, u_n is just a sequence of real numbers, but the story would be the same if it were a sequence of random variables. The point is that even if one random variable is well-approximated by another, that approximation may fail when conditioning on extremely rare events. It is rare though not impossible to observe an extremely negative value for M_t, which may cause problems for the Gaussian approximation deep in the left tail. Thus, we address this by essentially limiting full use of the truncation information, e.g. clipping the critical value C_t at 0. It is possible that a more careful analysis could relax this clipping.

Empirically, however, we do not observe significant loss in exactness from this choice (Section 5.1). We should not expect to, either, as at the boundary of the global null we should observe $C_t = 0$ only very rarely. This is by design, since the point of the correction is

to address the failure of the Gaussian approximation under rare events. Another way to see this is by noticing, from Equation (9), that M_t is highly negative only when $(Y_i)_{i \notin \mathcal{R}_t}$ contains substantially many positive values. Under the global null that $\mathbb{E}[Y \mid X] = 0$ almost surely, it would be rare for chiseling to consecutively reveal a substantial quantity of positive Y as chiseling can only reveal points based on X which is not predictive of the mean of Y.

Lastly, we note that while technically the asymptotic subtlety could be addressed by clipping C_t at any value, the clipping of C_t at zero could be considered a methodological advantage, as it prevents us from rejecting $H_0 : \mu(\mathcal{R}_t) \leq 0$ when $\bar{\mu}_t$ is negative, which would result in the unusual (though not necessarily invalid from a frequentist standpoint) conclusion that the population region mean is significantly positive while the empirical region mean is negative. Again, however, this would only happen extremely rarely even in the absence of any clipping.

C.11 Two team cross-screening

Roy et al. (2025) propose a procedure called "two team cross-screening," which is like data splitting except that instead of splitting the dataset randomly, it is split based on some meaningful covariate. For instance, $\mathcal{D}_1$ may consist of all units below the age of 45 while $\mathcal{D}_2$ may consist of all units at or above the age of 45. $\mathcal{D}_1$ is used to plan an analysis for $\mathcal{D}_2$, and vice versa; if a hypothesis is selected in both planning phases and rejected in both datasets, then the finding is deemed highly credible.

This idea can be used to initialize chiseling without auxiliary randomness. We first observe that splitting on a meaningful covariate partitions $\mathcal{X}$ into disjoint sets $\mathcal{X}_1$ and $\mathcal{X}_2$, where the data in $\mathcal{D}_1$ is drawn from the former and the data in $\mathcal{D}_2$ is drawn from the latter. Next, we note that if $\mathcal{R}_1 \subseteq \mathcal{X}_1$ and $\mathcal{R}_2 \subseteq \mathcal{X}_2$, then $\mu(\mathcal{R}_1 \cup \mathcal{R}_2)$ is a convex combination of $\mu(\mathcal{R}_1)$ and $\mu(\mathcal{R}_2)$, and hence

$$\mu(\mathcal{R}_1) > 0 \text{ and } \mu(\mathcal{R}_2) > 0 \implies \mu(\mathcal{R}_1 \cup \mathcal{R}_2) > 0.$$

This holds even when $\mathcal{R}_1$ or $\mathcal{R}_2$ are equal to $\emptyset$, where we recall that $\mu(\emptyset) > 0$ is a true proposition. Thus, an idea is to use $\mathcal{D}_2$ to initialize chiseling to be run on $\mathcal{D}_1$ with level $\alpha/2$, producing $\mathcal{R}_1 \subseteq \mathcal{X}_1$, and vice versa to produce $\mathcal{R}_2 \subseteq \mathcal{X}_2$. Then

$$\mathbb{P}(\mu(\mathcal{R}_1) \leq 0 \text{ or } \mu(\mathcal{R}_2) \leq 0) \leq \mathbb{P}(\mu(\mathcal{R}_1) \leq 0) + \mathbb{P}(\mu(\mathcal{R}_2) \leq 0) \leq \alpha/2 + \alpha/2 = \alpha.$$

So reporting the union of the regions $\mathcal{R}_1 \cup \mathcal{R}_2 \subseteq \mathcal{X}$ controls Type I error.

This works best when the covariate being split on does not have an interacting effect with the other covariates, so that learned relationships in $\mathcal{D}_1$ can be largely transported to $\mathcal{D}_2$ and vice versa. Even when non-interaction is satisfied, however, this initialization can be less powerful than randomly initializing a single instance of chiseling due to the reduction of the overall significance levels for each instance of chiseling to $\alpha/2$, as well as the reduction in sample size for each instance.

C.12 Combining chiseling across different initial split proportions

In Section 4.1, we described Bonferroni aggregation, which combines the results of chiseling (or any method that relies on an initial split proportion p) across $p_1 \leq ... \leq p_k$ by simply
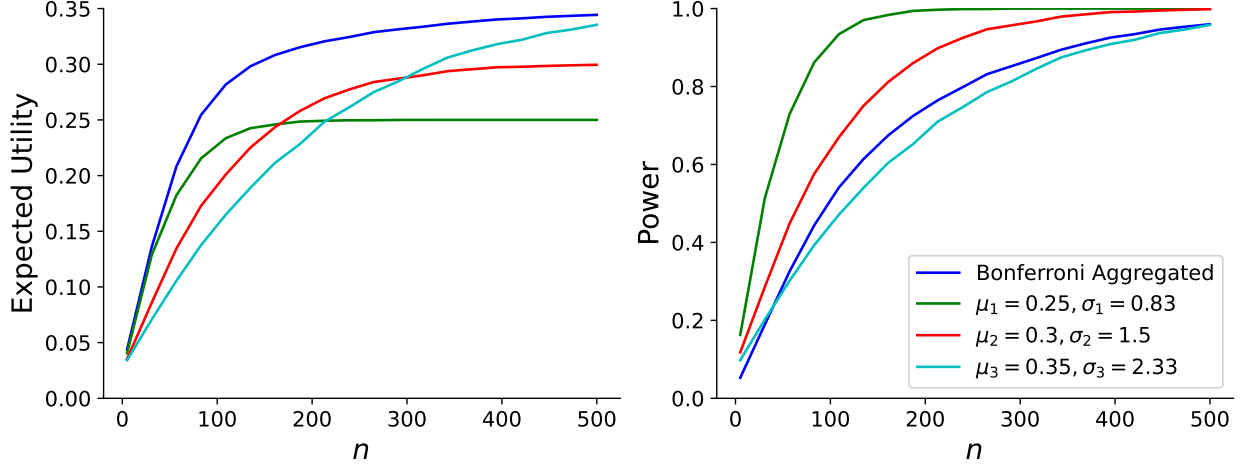
Figure C14: Stylized empirical demonstration of the effectiveness of Bonferroni aggregation.

running the analyses at adjusted levels $\alpha^{(1)}+...+\alpha^{(k)} = \alpha$. First, we make a remark regarding implementation: a valid but less effective way to realize Bonferroni combination is as follows. Consider, for instance, data splitting. Randomly shuffle and order the data points, and use the first p_1 proportion to train and the last $1 - p_1$ proportion to test, then use the first p_2 proportion to train and the last $1 - p_2$ proportion to test, and so on, finally aggregating the results of these k procedures via Bonferroni combination. This is clearly valid, but it is ineffective because there is high positive correlation between the different analyses, which makes the Bonferroni correction conservative. However, if we reshuffle the data before *each* analysis so that the $p_j \cdot n$ training points for the jth analysis are not simply a subset of the $p_{j+1} \cdot n$ training points for the $(j + 1)$st analysis, this makes the analyses more independent and alleviates some of Bonferroni's conservativeness. This is what we do in our simulations and we recommend it in practice.

Next, we discuss why Bonferroni aggregation is effective. Consider data splitting for simplicity. Typically, the utility of the tested subgroup will be higher for p_2-data splitting than for p_1-data splitting if $p_1 < p_2$. However, the power may sometimes (but not always) be lower for p_2 because the testing sample size is smaller, which brings down the expected utility. Suppose we have $p_1 < p_2 < p_3$. We can stylize the behavior of data splitting as follows: let $\mu_1 < \mu_2 < \mu_3$ and $\mu_1/\sigma_1 > \mu_2/\sigma_2 > \mu_3/\sigma_3$. The μ_j represent the mean of the tested subgroups discovered by p_j-data splitting, which are increasing in j but such that the signal size μ_j/σ_j is decreasing in j. We let $Y_{ij} \sim \mathcal{N}(\mu_j, \sigma_j^2)$ independently for $i = 1, ..., n$ and $j = 1, 2, 3$. The individual analyses test $H_j : \mu_j \leq 0$ at level α using the z-test and receive utility μ_j if a rejection is made and 0 otherwise. The Bonferroni aggregated analysis tests H_1, H_2, H_3 simultaneously at levels $\alpha/3$, then receives utility μ_j for the largest j such that H_j rejected and 0 if no rejection is made.

We can understand the behavior in a limiting setting where $\alpha \to 0$ and the signal sizes are scaled so that the testing problem is non-trivial. Let $R_j(\alpha)$ denote the indicator of the event that we can reject H_j at level α and let $q_j(\alpha) = \mathbb{E}[R_j(\alpha)]$. As $\alpha \to 0$, the worst-case power loss between running the z-test at level α and level $\alpha/3$ converges to 0 (see Appendix C.13 for a related discussion). This means that for α in a sufficiently small neighborhood of 0, we

have that the expected utility of Bonferroni aggregation is

$$\mathbb{E}[\max\{\mu_1 R_1(\alpha/3), \mu_2 R_2(\alpha/3), \mu_3 R_3(\alpha/3)\}] \geq \mathbb{E}[\mu_j R_j(\alpha/3)] = \mu_j q_j(\alpha/3) \approx \mu_j q_j(\alpha)$$

for all $j = 1, 2, 3$. Thus, the expected utility of Bonferroni aggregation is no worse than that of any individual test when α is sufficiently small. In fact, it can be much larger, since the mean of the maximum is generally larger than the maximum of the individual means. We verify in a toy example that even when $\alpha = 0.05$ the difference can be quite stark (Figure C14).

In practice, the improvements we observe are not as dramatic because the tests are not independent but positively correlated. However, we do find Bonferroni aggregation to be surprisingly effective and we believe this stylized example hints at why.

C.13 Weak non-inferiority of chiseling to the global t-test

Suppose without loss of generality that $\mathrm{Var}(Y) = 1$ and $\mu_{\mathrm{cut}} = 0$. Under local asymptotic scaling where $\mu(\mathcal{X}) = \tau/\sqrt{n}$, the asymptotic power of the t-test for testing $H_0 : \mu(\mathcal{X}) \leq 0$ at level α is given by $\Phi(\tau - z_{1-\alpha})$ where $z_{1-\alpha}$ is the $1 - \alpha$ quantile of a standard normal (see, for instance, Spiess et al. (2023)). Then if chiseling sets $\alpha_0 = c\alpha$ for $c \in [0, 1]$ the asymptotic difference in power between chiseling and the t-test can be no greater than

$$L(c; \alpha) := \sup_{\tau \in \mathbb{R}} \Phi(\tau - z_{1-\alpha}) - \Phi(\tau - z_{1-c\alpha}).$$

For instance, $L(0.5; 0.05) \approx 0.125$, while $L(0.5; 0.0001) \approx 0.068$. $L(c; \alpha)$ is decreasing in c and increasing in α (see the end of this subsection). This is the worst case loss in power, attained when the effect is completely homogeneous, but of course chiseling can achieve much greater power if the remaining $(1 - c)\alpha$ of the error can be effectively utilized to learn a subgroup, as we have seen in Section 5.5. The analyst may resort to curves such as those in Figure C15 in order to select an acceptable level of non-inferiority relative to the t-test, trading this off with the strength of their prior confidence in being able to detect subgroups.

When $\mathbb{E}[Y \mid X] \geq 0$ almost surely, then this bound on the power difference immediately translates to a bound on the difference in normalized expected utility. Let A be the event that chiseling rejects $\mathcal{R}_0 = \mathcal{X}$ and let B be the event that the global t-test rejects. Then

$$\mathbb{E}[\mathcal{U}(\mathcal{R}_\tau)] = \mathbb{E}[\mathcal{U}(\mathcal{R}_\tau) \mid A]\mathbb{P}(A) + \mathbb{E}[\mathcal{U}(\mathcal{R}_\tau) \mid A^c]\mathbb{P}(A^c) \geq \mathcal{U}(\mathcal{X})\mathbb{P}(A) = \mathcal{U}(\mathcal{R}^*)\mathbb{P}(A)$$

where $\mathcal{R}^*$ is an optimal utility region (because in this case $\mathcal{X}$ is an optimal utility region). The inequality follows from the fact that $\mathcal{U}(\mathcal{R}_\tau) \geq 0$ almost surely since $\mathbb{E}[Y \mid X] \geq 0$. Then as the expected utility of the t-test is $\mathcal{U}(\mathcal{R}^*)\mathbb{P}(B)$,

$$\frac{\mathcal{U}(\mathcal{R}^*)\mathbb{P}(B) - \mathbb{E}[\mathcal{U}(\mathcal{R}_\tau)]}{\mathcal{U}(\mathcal{R}^*)} \leq \frac{\mathcal{U}(\mathcal{R}^*)\mathbb{P}(B) - \mathcal{U}(\mathcal{R}^*)\mathbb{P}(A)}{\mathcal{U}(\mathcal{R}^*)} = \mathbb{P}(B) - \mathbb{P}(A) \leq L(c; \alpha).$$

Proof that $L(c; \alpha)$ is increasing in c and decreasing in α. Fix c and α. It suffices to consider $c \in (0, 1)$. Let ϕ be the density of a standard normal. Setting the derivative of $\Phi(\tau - z_{1-\alpha}) - \Phi(\tau - z_{1-c\alpha})$ with respect to τ equal to 0, we get

$$\phi(\tau - z_{1-\alpha}) - \phi(\tau - z_{1-c\alpha}) = 0.$$
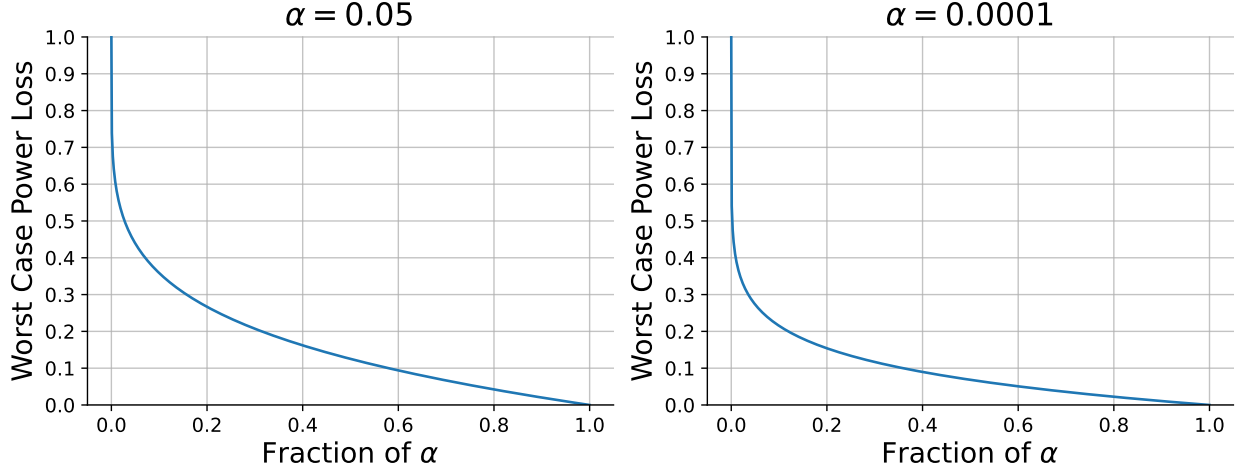
Figure C15: Asymptotic worst case power loss of running a t-test with a fraction of α compared to running the t-test at level α.

Since ϕ is symmetric and strictly decreasing away from 0, and since $z_{1-\alpha} \neq z_{1-c\alpha}$, this holds if and only if $\tau - z_{1-\alpha} = -(\tau - z_{1-c\alpha})$, i.e. $\tau = \frac{z_{1-\alpha}+z_{1-c\alpha}}{2}$. So we have

$$L(c;\alpha) = \Phi\left(\frac{z_{1-c\alpha} - z_{1-\alpha}}{2}\right) - \Phi\left(\frac{-z_{1-c\alpha} + z_{1-\alpha}}{2}\right) = 2\Phi\left(\frac{z_{1-c\alpha} - z_{1-\alpha}}{2}\right) - 1.$$

Since Φ is increasing, it suffices to show that $z_{1-c\alpha} - z_{1-\alpha}$ is decreasing in c and increasing in α. The former is easy to see, so we focus on the latter. By the inverse function rule, the derivative of this quantity with respect to α is given by

$$\frac{-c}{\phi(z_{1-c\alpha})} + \frac{1}{\phi(z_{1-\alpha})} = \frac{1}{\alpha}\left(\frac{1 - \Phi(z_{1-\alpha})}{\phi(z_{1-\alpha})} - \frac{1 - \Phi(z_{1-c\alpha})}{\phi(z_{1-c\alpha})}\right).$$

Since $z_{1-c\alpha} > z_{1-\alpha}$, it suffices to show that $\frac{1-\Phi(x)}{\phi(x)}$ is decreasing in x. But this is simply the Mills ratio, and it is a well-known fact that this quantity is strictly decreasing on $\mathbb{R}$ (see, for instance, Baricz (2008)). $\qquad\square$

C.14   Shrinking along hyperrectangular contours

Let $f : \mathbb{R}^d \to \mathbb{R}$ be any function. Define $A_z := \{x \in \mathbb{R}^d : f(x) > z\}$. We show that A_z is hyperrectangular for all z if and only if $f(x) = \min_{j=1,\dots,d}\{f_j(x^{(j)})\}$ for some univariate functions f_j whose upper level sets are intervals. The reverse direction was shown in Section 4.5, so we focus on the forward direction. By assumption, $A_z = \prod_{j=1}^d I_j(z)$ where $I_j(z)$ are intervals such that $I_j(z) \supseteq I_j(z')$ for all $z \leq z'$. Define $f_j(x^{(j)}) = \sup\{z \in \mathbb{R} : x^{(j)} \in I_j(z)\}$. Then

$$x^{(j)} \in \{x' : f_j(x') > z\} \iff f_j(x^{(j)}) > z \iff \exists z' > z \text{ s.t. } x^{(j)} \in I_j(z') \implies x^{(j)} \in I_j(z).$$

To reverse the last implication, suppose without loss of generality that $j = 1$. Note that $x^{(1)} \in I_1(z)$ implies that there exists $u \in \mathbb{R}^{d-1}$ such that $f(x^{(1)}, u) > z$, and thus there exists

z' such that $f(x^{(1)}, u) > z' > z$. But this implies that $x^{(1)} \in I_1(z')$, so the last implication may be reversed, and we have that the upper level sets of $f_j(\cdot)$ are intervals. Then finally note that

$$f(x) > z \iff \forall j \ x^{(j)} \in I_j(z) \iff \forall j \ f_j(x^{(j)}) > z \iff \min_{j=1,..,d} f_j(x^{(j)}) > z.$$

The upper level sets of $f(\cdot)$ and $\min_{j=1,...,d} f_j(x^{(j)})$ are identical for all z and hence the two functions are equal. This last fact can be quickly seen by contradiction: for functions f and g, suppose the upper level sets of f and g agree but $f(x') < g(x')$ for some x'. Then there is a z such that $f(x') < z < g(x')$, but then $x' \notin \{x : f(x) > z\}$ while $x' \in \{x : g(x) > z\}$, a contradiction.

C.15 Rejecting multiple regions

We show that the chiseling procedure we have developed can be used to reject multiple regions while controlling the family-wise error rate (FWER) as long as it is used in an appropriate way. The construction is essentially an application of the closed testing principle. There are, however, some subtleties which arise from the adaptive nature of our problem.

Imagine the following procedure: the analyst proceeds through the chiseling algorithm as usual, producing a sequence of regions $\mathcal{R}_0, ..., \mathcal{R}_m$. Let us suppose that the analyst does not run any tests (and hence does not terminate until after the last region $\mathcal{R}_m$ has been defined). Instead, at stage t, the analyst picks a value $v_t(S) \in [0, 1]$ for every $S \subseteq [t]_0$ such that $t \in S$, and sets $v_t(S) = 0$ for every other $S \subseteq [t]_0$, We require that the choices satisfy $\prod_{s \leq t} 1 - v_s(S \cap [s]_0) \geq 1 - \alpha$ for all such S. We call $v_t(S)$ the *hypothetical α budgets*. After defining all the hypothetical α budgets, we can go back and do the following: for each stage t, let $\phi_t = 1$ if, for every $S \subseteq [t]_0$ such that $t \in S$, chiseling run with $\alpha_s = v_s(S \cap [s]_0)$ for $s \leq t$ rejects *some* region $\mathcal{R}_s$ with $s \leq t$. Then

$$\mathbb{P}(\exists t \text{ s.t. } \phi_t = 1 \text{ and } \mu(\mathcal{R}_t) \leq \mu_{\text{cut}}) \leq \alpha$$

when the tests are finite-sample valid, and

$$\limsup_{n \to \infty} \mathbb{P}(\exists t \text{ s.t. } \phi_t = 1 \text{ and } \mu(\mathcal{R}_t) \leq \mu_{\text{cut}}) \leq \alpha$$

when the tests are asymptotically valid. We prove this below and then sketch out a simple and illuminating example. In general this procedure can be computationally intensive for the same reasons that closed testing is computationally intensive (i.e. needing to test every intersection null), but computational shortcuts are possible if structure is imposed on $v_t(S)$.

Proof. The analyst defines, in the course of this modified variant of chiseling, a set of hypothetical α budgets. We will consider an oracle that runs the usual variant of chiseling by picking out and abiding by a single hypothetical α budget corresponding to testing all the true nulls. The oracle knows $\mathcal{N}_t := \{s \leq t : \mu(\mathcal{R}_s) \leq \mu_{\text{cut}}\} \subseteq [t]_0$ at stage t since $\mathcal{R}_s$ is $\mathcal{F}_t$-measurable for $s \leq t$. We let the oracle set $\alpha_t = v_t(\mathcal{N}_t)$. By construction,

$$\prod_{t=0}^{m} 1 - \alpha_t = \prod_{t=0}^{m} 1 - v_t(\mathcal{N}_t) = \prod_{t=0}^{m} 1 - v_t(\mathcal{N}_m \cap [t]_0) \geq 1 - \alpha$$

so this is a valid α-budget. Now we define the oracle multiple rejection set $\mathcal{V}_{\text{oracle}} := \mathcal{N}_m^{\mathsf{c}}$ if the oracle instance of chiseling does not reject (i.e. reports the empty region), and $\mathcal{V}_{\text{oracle}} := [m]_0$ otherwise. In words, the oracle rejects all the false nulls (and makes no errors) if its instance of chiseling accepts, and otherwise rejects every hypothesis. Hence, the oracle FWER is bounded by the probability that the oracle instance of chiseling rejects. Since the oracle is only testing null regions and chiseling is valid, the probability of this event is no greater than α, and thus the oracle FWER is bounded by α.

Now we argue that $\mathcal{V} := \{t : \phi_t = 1\} \subseteq \mathcal{V}_{\text{oracle}}$ deterministically. Suppose $t \in \mathcal{V}$. If $\mu(\mathcal{R}_t) > \mu_{\text{cut}}$, then $t \in \mathcal{V}_{\text{oracle}}$ since $\mathcal{V}_{\text{oracle}}$ always contains all the false nulls. If $\mu(\mathcal{R}_t) \leq \mu_{\text{cut}}$, then the oracle instance of chiseling must have rejected at some stage $s \leq t$. This is because $t \in \mathcal{N}_t$ and the oracle α-budget up to stage t corresponds to $(v_0(\mathcal{N}_0), ..., v_t(\mathcal{N}_t)) = (v_0(\mathcal{N}_t \cap [0]_0), ..., v_t(\mathcal{N}_t \cap [t]_0))$, but every hypothetical instance of chiseling that uses α-budget $(v_0(S \cap [0]_0), ..., v_t(S \cap [t]_0))$ rejects for any $S \subseteq [t]_0$ as long as $t \in S$; this is by definition of t belonging to $\mathcal{V}$. Then $\mathcal{V}_{\text{oracle}} = [m]_0$ and trivially includes t. Thus,

$$\mathbb{P}(\exists t \in \mathcal{V} \text{ s.t. } \mu(\mathcal{R}_t) \leq \mu_{\text{cut}}) \leq \mathbb{P}(\exists t \in \mathcal{V}_{\text{oracle}} \text{ s.t. } \mu(\mathcal{R}_t) \leq \mu_{\text{cut}}) \leq \alpha$$

when the tests are finite-sample valid. The first inequality is always true regardless of the validity of the chiseling. Thus, when the tests are asymptotic we can simply put a $\limsup$ in front of the probabilities. $\qquad\square$

Example: rejecting two regions. Suppose we wish to test $\mathcal{R}_0 = \mathcal{X}$ and one additional chiseled region. At a high level, the closed testing construction allows us to do following: we pick some α_0, and then use chiseling to shrink and stop at one region $\mathcal{R}_\nu$ (e.g. Algorithm 3). If we do not reject $H_0 : \mu(\mathcal{R}_0) \leq \mu_{\text{cut}}$ at level α_0, we apply the sequential testing framework developed in this paper and test $H_\nu : \mu(\mathcal{R}_\nu) \leq \mu_{\text{cut}}$ at level $\alpha_\nu = 1 - (1 - \alpha)/(1 - \alpha_0)$ conditionally on having accepted the previous test, rejecting if possible. If we do reject H_0, we report that, and then we are entitled to test H_ν *unconditionally* at level α and also report that if it rejects. The intuition is that when we do reject H_0, there are only two scenarios: H_0 is null, in which case we have already committed a Type I error and additional rejections are "free," or (2) H_0 is not null, in which case there was no need to test it and we could have, in the first place, saved all of our α for testing H_ν. Note that here we have implicitly set, for every $t \notin \{0, \nu\}$, $v_t(S) = 0$ for every $S \subseteq [t]_0$, which renders the test feasible to execute.

C.16 Point estimates and confidence intervals

Beyond testing, a standard analysis would provide point estimates and confidence intervals.

Point estimates. $\bar{\mu}_t = \frac{1}{n_t} \sum_{i:X_i \in \mathcal{R}_t} Y_i$ is the natural estimate for $\mu(\mathcal{R}_t)$. In fact, Corollary 3.3 guarantees that $\bar{\mu}_t$ is unbiased for $\mu(\mathcal{R}_t)$ conditionally on $\mathcal{F}_t$ as long as there is at least one point in $\mathcal{R}_t$. If we take the convention that an undefined value—call it NaN—is equal to any number, i.e. $\text{NaN} - x = 0$ for all x, then by applying the tower rule to each coordinate we have

$$\mathbb{E}[(\bar{\mu}_0, ..., \bar{\mu}_m) - (\mu(\mathcal{R}_0), ..., \mu(\mathcal{R}_m))] = (0, ..., 0).$$

Though the convention that $\text{NaN} - x = 0$ for all x appears odd, it is not artificial. It simply reflects the idea that if we have no estimate to report, then we are neither overestimating nor underestimating the truth. It is thus appropriate to call NaN a trivially unbiased estimate. We note, of course, that the above unbiasedness does not hold *conditionally* on having made a rejection, for instance using Algorithm 1. But neither is any ordinarily unbiased point estimate $\hat{\theta}$ for θ conditional on having either accepted or rejected $H_0 : \theta \leq 0$, say, using a t-test. Thus, it may not be unreasonable to interpret $\bar{\mu}_t$ similarly to how one would interpret an unbiased estimate in the context of a larger analysis that includes hypothesis tests (i.e. as unbiased prior to testing).

In fact, the natural unbiased estimates have some reasonable consistency properties. For example, suppose that $p \in (0, 1)$, Y is bounded, and $m = o(e^{pn})$.[8] Then

$$\lim_{n \to \infty} \mathbb{P}\left(\sup_{t: n_t \geq pn} |\bar{\mu}_t - \mu(\mathcal{R}_t)| > \epsilon \right) = 0 \quad \text{for all } \epsilon > 0.$$

The result follows from applying Hoeffding's inequality conditionally. For some constant C, the following holds for all t:

$$\mathbb{P}(|\bar{\mu}_t - \mu(\mathcal{R}_t)| > \epsilon \mid \mathcal{F}_t) \cdot \mathbb{1}\{n_t \geq pn\} \leq Ce^{-n_t} \cdot \mathbb{1}\{n_t \geq pn\} \leq Ce^{-pn} \quad \text{a.s.}$$

Then

$$\mathbb{P}\left(\sup_{t: n_t \geq pn} |\bar{\mu}_t - \mu(\mathcal{R}_t)| > \epsilon \right) \leq \sum_{t=0}^{m} \mathbb{P}(|\bar{\mu}_t - \mu(\mathcal{R}_t)| > \epsilon \text{ and } n_t \geq pn) \leq C(m+1)e^{-pn} \to 0$$

by observing that the summands in the second expression are simply the expectations of the left-hand side of the previous display. It is probably possible to weaken the condition that Y is bounded, but this requires more technical effort.

Algorithm 4 Generic multiple testing via chiseling

    Input: dataset $\mathcal{D} = (X_i, Y_i)_{i=1}^{n}$ and nominal family-wise error rate α

1: Initialize region $\mathcal{R}_0 \leftarrow \mathcal{X}$ and revealed information $\mathcal{F}_0 \leftarrow \{\emptyset, \Omega\}$
2: **for** $t = 0, ..., m$ **do**
3:     Choose a hypothesis $\mathcal{H}_t$ and $\alpha_t \leq \alpha - \sum_{s=0}^{t-1} \alpha_s$ based on $\mathcal{F}_t$
4:     Perform a valid level α_t test ϕ_t of $\mathcal{H}_t$ conditional on $\mathcal{F}_t$
5:     Obtain $\mathcal{R}_{t+1} \subseteq \mathcal{R}_t$ by chiseling $\mathcal{R}_t$ based on $\mathcal{F}_t$ using Algorithm 2
6:     See revealed data points: $\mathcal{F}_{t+1} \leftarrow \sigma(\mathcal{F}_t, (X_i, Y_i)_{i: X_i \in \mathcal{R}_t \setminus \mathcal{R}_{t+1}})$
7: **end for**
8: **Return:** hypotheses $(\mathcal{H}_t)_{t=0}^{m}$ and rejections $(\phi_t)_{t=0}^{m}$

Confidence intervals. First, Algorithm 4 describes a generic multiple testing wrapper for chiseling based on the Bonferroni correction. Note that in the final line we return the

---

hypotheses $(\mathcal{H}_t)_{t=0}^m$ but here we should think of the hypotheses as being the adaptively determined null propositions about the data that are being tested by the analyst, not the actual values of $(\mathcal{H}_t)_{t=0}^m$ which are of course unobservable. The family-wise error control of Algorithm 4 follows directly from the union bound:

$$\mathbb{P}(\exists t \text{ s.t. } \mathcal{H}_t = 0 \text{ and } \phi_t = 1) \leq \sum_{t=0}^m \mathbb{P}(\mathcal{H}_t = 0, \phi_t = 1) = \sum_{t=0}^m \mathbb{E}[\mathbb{P}(\mathcal{H}_t = 0, \phi_t = 1 \mid \mathcal{F}_t)]$$

$$\leq \sum_{t=0}^m \mathbb{E}[\alpha_t] = \mathbb{E}\left[\sum_{t=0}^m \alpha_t\right] \leq \mathbb{E}[\alpha] = \alpha.$$

We clarify Line 4 of Algorithm 4. This is different from Definition 3.6 which requires validity conditional on the outcomes of the previous tests $(\phi_s)_{s<t}$. Here, we only require validity conditional on $\mathcal{F}_t$. Generally, we can achieve this with little fuss using Corollary 3.3. For instance, imagine if $\mathcal{H}_t = \mathbb{1}\{\text{median}(\mathcal{R}_t) \neq 0\}$ (i.e. $\mathcal{H}_t : \text{median}(\mathcal{R}_t) = 0$) where $\text{median}(\mathcal{R})$ is the median of $Y \mid X \in \mathcal{R}$. Then for every t we may simply let ϕ_t be the sign test applied to the samples $(Y_i)_{i:X_i \in \mathcal{R}_t}$ with level α_t.

Now we note that we may replace $\mathcal{H}_t$ with an estimand θ_t, ϕ_t with a $1 - \alpha_t$ confidence interval $\mathcal{C}_t$ for θ_t that is valid conditional on $\mathcal{F}_t$, and the return statement with the collection of selected estimands and confidence intervals. Then the same union bound argument implies

$$\mathbb{P}(\exists t \text{ s.t. } \theta_t \notin \mathcal{C}_t) \leq \alpha.$$

In other words, we have simultaneous coverage. To give a concrete example: we may let $\theta_t := \text{median}(\mathcal{R}_t)$ and $\mathcal{C}_t$ be the $1-\alpha_t$ confidence interval for the median obtained by inverting the sign test using the samples $(Y_i)_{i:X_i \in \mathcal{R}_t}$.

Note, of course, that this may be much less powerful than what we describe in the main text because we are not accounting for the dependence among tests. However, it is more general, precisely because it is valid under the worst-case dependence structure. What is lost in power is gained here in flexibility.

Lastly, if we replace Line 4 with "construct a valid level-α confidence interval $\mathcal{C}_t$ for θ_t conditional on $\mathcal{F}_t$," then we control the expected miscoverage rate:

$$\mathbb{E}\left[\frac{1}{m+1}\sum_{t=0}^m \mathbb{1}\{\theta_t \notin \mathcal{C}_t\}\right] \leq \alpha.$$

This is the analogue of providing pointwise coverage. For instance, if $\theta_t = \mu(\mathcal{R}_t)$ and we set aside asymptotics for the moment, then $(\mathcal{C}_0, ..., \mathcal{C}_m)$ is precisely like a pointwise confidence band for the process traced out by $(\mu(\mathcal{R}_0), ..., \mu(\mathcal{R}_m))$. While pointwise coverage is a weak statistical guarantee, it can be useful in exploratory contexts.

C.17 Finite-sample approach for binary potential outcomes

Suppose that our outcomes Y are binary, that we additionally measure a binary covariate W, and that we are interested in finding a region $\mathcal{R}$ such that there is a positive contrast, i.e. that $\mathbb{E}[Y \mid W = 1, X \in \mathcal{R}] > \mathbb{E}[Y \mid W = 0, X \in \mathcal{R}]$. If W is a vector of randomized

treatment assignments and Y is the outcome of the randomized experiment, then this is equivalent to a region where the average treatment effect is positive, i.e. this would reduce to $\mathbb{E}[Y(1) \mid X \in \mathcal{R}] > \mathbb{E}[Y(0) \mid X \in \mathcal{R}]$. Alternatively, since the outcomes are binary and the subgroup means are rates, it may be natural to desire regions such that the subgroup rate ratio is at least some value, i.e.

$$\frac{\mathbb{E}[Y \mid W = 1, X \in \mathcal{R}]}{\mathbb{E}[Y \mid W = 0, X \in \mathcal{R}]} > \text{ some fixed threshold.}$$

We will see that the ideas presented here can be readily extended to testing properties of arbitrary functions of $\mathbb{E}[Y \mid W = 1, X \in \mathcal{R}]$ and $\mathbb{E}[Y \mid W = 0, X \in \mathcal{R}]$, but for simplicity we will focus on the case of positive contrasts.

Let us use the notation $\mu_j(\mathcal{R}) = \mathbb{E}[Y \mid W = j, X \in \mathcal{R}]$ for $j = 0, 1$. Our situation would be very similar to the binary regression setting of Section 3.6 if not for the appearance of nuisance parameters. Note that even in the non-adaptive setting where $\mathcal{R}$ is fixed a priori, there are nuisances in testing $H_0 : \mu_1(\mathcal{R}) \leq \mu_0(\mathcal{R})$ as we care about the single unknown contrast but not the two unknown means. For a general test statistic T, it becomes necessary to search over the $1 - \alpha$ quantiles of T under all nulls that are consistent with H_0 in order to properly calibrate the null distribution. This entails a two-dimensional grid search. The goal of this section is to illustrate that this grid search is readily incorporated into our adaptive testing framework, where now the algorithm must perform a grid search at each stage where the analyst wishes to perform a test.

Define $n_{t,1} = \sum_{i=1}^n W_i \mathbb{1}\{X_i \in \mathcal{R}\}$ and $n_{t,0} = \sum_{i=1}^n (1 - W_i)\mathbb{1}\{X_i \in \mathcal{R}\}$, and suppose that the analyst is shown $n_{0,1}$ and $n_{0,0}$ at the outset, i.e. we condition on the sample sizes of the treated and control groups, and let $\mathcal{F}_0 = \sigma(n_{0,1}, n_{0,0})$. Then $n_{t,1}$ and $n_{t,0}$ are $\mathcal{F}_t$-measurable. Let us suppose that at stage t the analyst selects, using the information in $\mathcal{F}_t$, a test statistic T_t, which by abuse of notation we will suppose is both a function from $[n_{t,1}]_0 \times [n_{t,0}]_0$ to $\mathbb{R}$, and also the value it attains when applied to the following statistics of the masked data,

$$T_t := T_t\left(\sum_{i=1}^n W_i Y_i \mathbb{1}\{X_i \in \mathcal{R}_t\}, \sum_{i=1}^n (1 - W_i)Y_i \mathbb{1}\{X_i \in \mathcal{R}_t\}\right). \tag{10}$$

Note that the actual values that T_t outputs do not matter so much as the ranking it induces on the elements of $[n_{t,1}]_0 \times [n_{t,0}]_0$. Ultimately, rejecting for large values of T_t will carve out a subset of $[n_{t,1}]_0 \times [n_{t,0}]_0$ corresponding to rejections, and *not* rejecting will tell us that the above statistics of the masked data are *not* in this subset. We can show that this information propagates cleanly from stage to stage. At stage t, we recursively define the *allowed pairs* Ξ_t and critical values C_t at stage t via the following procedure. Begin with $\Xi_0 := [n_{0,1}]_0 \times [n_{0,0}]_0$, and recursively define the following for $t = 0, ..., m$:

$$C_t := \sup_{(p_1, p_0) \in [0,1]^2 : p_1 \leq p_0} Q_{\mathrm{RB}}(1 - \alpha_t; (n_{t,1}, n_{t,0}), (p_1, p_0), \Xi_t)(T_t),$$

$$\Delta_{t,1} := \sum_{i=1}^n W_i Y_i \mathbb{1}\{X_i \in \mathcal{R}_t \setminus \mathcal{R}_{t+1}\},$$

$$\Delta_{t,0} := \sum_{i=1}^n (1 - W_i)Y_i \mathbb{1}\{X_i \in \mathcal{R}_t \setminus \mathcal{R}_{t+1}\},$$

$$\Xi_{t+1} := \{(j - \Delta_{t,1}, k - \Delta_{t,0}) : (j, k) \in \Xi_t \text{ and } T_t(j, k) \leq C_t\} \cap ([n_{t+1,1}]_0 \times [n_{t+1,0}]_0)$$

<div style="text-align:center">(11)</div>

where, for a function $f : [n_1]_0 \times [n_0]_0 \to \mathbb{R}$, we define the *restricted binomial statistic quantile* $Q_{\mathrm{RB}}(q; (n_1, n_0), (p_1, p_0), \Xi)(f)$ as

$$Q_{\mathrm{RB}}(q; (n_1, n_0), (p_1, p_0), \Xi)(f) := \inf\{z \in \mathbb{R} : \mathbb{P}(f(B_1, B_2) \le z \mid (B_1, B_2) \in \Xi) \ge q\},$$

where B_1 and B_2 are drawn independently from $\mathrm{Binom}(n_1, p_1)$ and $\mathrm{Binom}(n_0, p_0)$, respectively. It is possible to allow randomized versions of Q_{RB} as in Appendix C.8, but we will not discuss this.

The complexity of evaluating $Q_{\mathrm{RB}}(q; (n_1, n_0), (p_1, p_0), \Xi)(f)$ is essentially $O(n_1 \cdot n_0)$, assuming f takes constant time to evaluate. Though the additional search over (p_1, p_0) makes evaluating C_t more expensive, it is still within the capabilities of modern personal computers, especially when the sample sizes are modest. For $n_1 = n_0 = 100$, it takes a matter of seconds to calculate $Q_{\mathrm{RB}}(q; (n_1, n_0), (p_1, p_0), \Xi)(f)$ over 20,000 pairs of $(p_1, p_0) \in [0, 1]^2$ on a personal computer. There may be further computational improvements. For instance, it may be possible that for certain choices of test statistics, the solution to the optimization is always attained on the diagonal $p_1 = p_0$.

Though these definitions may appear convoluted, the motivation for these constructions is the same as for our other tests. As we show at the end of this section,

$$T_s \le C_s \text{ for } s = 1, ..., t-1 \iff \left(\sum_{i=1}^{n} W_i Y_i \mathbb{1}\{X_i \in \mathcal{R}_t\}, \sum_{i=1}^{n} (1 - W_i) Y_i \mathbb{1}\{X_i \in \mathcal{R}_t\} \right) \in \Xi_t. \tag{12}$$

Conditionally on $\mathcal{F}_t$, the pair on the right-hand side is distributed as $\mathrm{Binom}(n_{t,1}, p_1)$ in the first element and as $\mathrm{Binom}(n_{t,0}, p_0)$ in the second element, for some $p_1 \le p_0$ under the null. Knowing that we have not yet rejected simply reduces the support to Ξ_t, and this motivates the definition of Q_{RB}. The proof of the following theorem follows that of Appendix D.3.1 exactly except we replace truncation by the restriction of pairs given by Equation (12). As the proof is otherwise redundant, we omit it.

Theorem C.1 (Validity of test for binary contrasts). *Under the setting described in this section, let $\mathcal{R}_\tau$ be the region produced by Algorithm 1 using any test statistic of the form described by Equation (10), and using Equation (11) to compute the critical values. Then*

$$\mathbb{P}(\mu_1(\mathcal{R}_\tau) \le \mu_0(\mathcal{R}_\tau)) \le \alpha.$$

To implement this test, it is natural to let the test statistic T_t be the standardized difference in means, which can be calculated from $\sum_{i=1}^{n} W_i Y_i \mathbb{1}\{X_i \in \mathcal{R}_t\}$, $\sum_{i=1}^{n} (1 - W_i) Y_i \mathbb{1}\{X_i \in \mathcal{R}_t\}$ and the information in $\mathcal{F}_t$, though other statistics may also be reasonable. We leave an empirical exploration to future work.

Remark C.2 (Targeted unmasking). *Since the relative sample sizes $n_{t,1}$ and $n_{t,0}$ are ancillary to the parameters of interest in this setting, whenever the analyst wishes to reveal a point at random, she may opt to selectively reveal whether this point should be from the treated group or the control group (as opposed to revealing a point from either group at random).*

Proof of Equation (12). Define

$$Z_t^{(1)} := \sum_{i=1}^{n} W_i Y_i \mathbb{1}\{X_i \in \mathcal{R}_t\} \quad \text{and} \quad Z_t^{(0)} := \sum_{i=1}^{n} (1 - W_i) Y_i \mathbb{1}\{X_i \in \mathcal{R}_t\}.$$

We show the equivalence in Equation (12) by induction. The case $t = 0$ is vacuously true. Suppose it is true for $t - 1$. Then the inductive hypothesis yields

$$T_s \leq C_s \text{ for } s = 1, ..., t - 2 \iff (Z^{(1)}_{t-1}, Z^{(0)}_{t-1}) \in \Xi_{t-1}$$

which implies

$$T_s \leq C_s \text{ for } s = 1, ..., t - 1 \iff (Z^{(1)}_{t-1}, Z^{(0)}_{t-1}) \in \Xi_{t-1} \text{ and } T_{t-1} \leq C_{t-1}.$$

Thus, it suffices to show that the right-hand side of the above is equivalent to $(Z^{(1)}_t, Z^{(0)}_t) \in \Xi_t$, which is the right-hand side of Equation (12).

Recall that $T_{t-1} = T_{t-1}(Z^{(1)}_{t-1}, Z^{(0)}_{t-1})$, that $Z^{(1)}_{t-1} - \Delta_{t-1,1} = Z^{(1)}_t$, and that $Z^{(0)}_{t-1} - \Delta_{t-1,0} = Z^{(0)}_t$. Hence, the right-hand side of the above implies that $(Z^{(1)}_t, Z^{(0)}_t) = (j - \Delta_{t-1,1}, k - \Delta_{t-1,0})$ for some $(j, k) \in \Xi_{t-1}$ that also satisfies $T_{t-1}(j, k) \leq C_{t-1}$. Then since we additionally have $(Z^{(1)}_t, Z^{(0)}_t) \in ([n_{t,1}]_0 \times [n_{t,0}]_0)$ by construction, this yields

$$(Z^{(1)}_{t-1}, Z^{(0)}_{t-1}) \in \Xi_{t-1} \text{ and } T_{t-1} \leq C_{t-1} \implies (Z^{(1)}_t, Z^{(0)}_t) \in \Xi_t$$

by the definition of Ξ_t. To show the reverse implication, note that $(Z^{(1)}_t, Z^{(0)}_t) \in \Xi_t$ implies that $(Z^{(1)}_t, Z^{(0)}_t) = (j - \Delta_{t-1,1}, k - \Delta_{t-1,0})$ for some $(j, k) \in \Xi_{t-1}$ that also satisfies $T_{t-1}(j, k) \leq C_{t-1}$. But this uniquely identifies $j = Z^{(1)}_{t-1}$ and $k = Z^{(0)}_{t-1}$, so we get the reverse implication. This completes the proof by induction. □

Appendix D   Proofs of main results

First, we describe some additional notation, conventions, and basic facts that will be used throughout the proofs. For a sequence of random variables X_n, write $X_n \xrightarrow{p} \infty$ to mean that $\mathbb{P}(X_n > B) \to 1$ for every $B \in \mathbb{R}$. It will sometimes be useful for us to condition on both a σ-algebra and an event. We formalize what this means. Let X be a random variable with finite first moment, A an event, and $\mathcal{F}$ a σ-algebra. We use the notation

$$\mathbb{E}[X \mid \mathcal{F}, A] = \frac{\mathbb{E}[X \mathbb{1}\{A\} \mid \mathcal{F}]}{\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}]}$$

where we leave $0/0$ undefined. We note that $\mathbb{E}[X \mid \mathcal{F}, A]$ is $\mathcal{F}$-measurable, and that

$$\mathbb{E}[X \mid \mathcal{F}, A] = \mathbb{E}[X \mid \mathcal{F}, \mathbb{1}\{A\}] \quad \text{a.s. conditional on } A.$$

For completeness, the above fact is proved as Lemma F.2 in Appendix F. In particular, we note that $\mathbb{E}[X \mid \mathcal{F}, A]$ is well-defined and finite almost surely conditional on A. In our proofs, we will only ever consider the values of such quantities conditional on A, and hence our particular convention for $0/0$ will not figure into our arguments.

Let Z be a discrete random variable. In general, $\mathbb{E}[X \mid Z]$ and $\mathbb{E}[X \mid h(Z)]$ are not equal unless h is invertible. However, they are equal for almost all ω belonging to $\{\omega \in \Omega : h^{-1}(h(Z(\omega)))$ is a singleton$\}$. That is, whenever $h(Z)$ takes on a value from which Z can be perfectly deduced, the two are interchangeable. For completeness, this is proved as Lemma F.3 in Appendix F.

D.1 Properties of chiseling

D.1.1 Preliminary results

Though the following result is somewhat obvious, we offer a proof in Appendix F.2 for completeness. It simply states that conditional on the clipped minimum of an i.i.d. real-valued sample, the points that do not attain the clipped minimum are i.i.d. and have distributions truncated to be above the minimum.

Lemma D.1 (Distribution conditional on minimum). *Let $S_1, ..., S_m$ be i.i.d. real-valued random variables. Let $c \in \mathbb{R} \cup \{\infty\}$ be fixed and define $\gamma := c \wedge \min_{i=1,...,m} S_i$ and $\mathcal{I} := \{i : S_i > \gamma\}$. Then the distribution of $(S_i)_{i \in \mathcal{I}}$ conditional on γ and $\mathcal{I}$ is the same as $|\mathcal{I}|$ i.i.d. draws from the distribution of $S \mid S > \gamma$.*

Now we isolate the essential idea of the proof of Theorem 3.2 as a standalone lemma.

Lemma D.2 (General thresholding). *Let $(Z_1, ..., Z_m)$ be i.i.d. samples from any distribution taking values in $\mathcal{Z}$ and let Z be an additional independent copy from the same distribution. Let $c \in \mathbb{R} \cup \{\infty\}$ and $g : \mathcal{Z} \to \mathbb{R}$ be fixed. Define random variables $\gamma := c \wedge (\min_{i=1,...,m} g(Z_i))$ and $m' := |\{i : g(Z_i) > \gamma\}|$. Also define $\mathbf{Z}_{\text{out}} := (Z_i)_{i:g(Z_i) \leq \gamma}$ and $\mathbf{Z}_{\text{in}} := (Z_i)_{i:g(Z_i) > \gamma}$. Then γ and m' are $\sigma(\mathbf{Z}_{\text{out}})$-measurable, and the distribution of $\mathbf{Z}_{\text{in}}$ given $\mathbf{Z}_{\text{out}}$ is the same as m' i.i.d. draws from the distribution of $Z \mid g(Z) > \gamma$.*

Proof. By construction, γ is equal to g applied to any element of $\mathbf{Z}_{\text{out}}$ if it is non-empty, and equal to c otherwise. Also, m' is equal to m minus the number of elements in $\mathbf{Z}_{\text{out}}$. Hence, γ and m' are $\sigma(\mathbf{Z}_{\text{out}})$-measurable. Define the random set $\mathcal{I} := \{i : g(Z_i) > \gamma\}$. Note that, by a symmetry argument, conditional on $\mathbf{Z}_{\text{out}}$, $\mathcal{I}$ is uniformly distributed over subsets of $\{1, ..., m\}$ of size m' independently of $\mathbf{Z}_{\text{in}}$. Thus $\mathbf{Z}_{\text{in}}$ is independent of $\mathcal{I}$ given $\mathbf{Z}_{\text{out}}$.

Since γ and $\mathcal{I}$ are functions of $(Z_1, ..., Z_m)$, they have a distribution implied by the distribution of $(Z_1, ..., Z_m)$. We describe a way of sampling $(Z_1, ..., Z_m)$ that reveals the underlying conditional independence.

1. Jointly sample the pair $(\gamma, \mathcal{I})$ from its marginal distribution.

2. For $i \notin \mathcal{I}$, set $G_i = \gamma$. Then independently sample Z_i from the distribution of $Z \mid g(Z) = G_i$

3. For $i \in \mathcal{I}$, independently sample G_i from the distribution of $g(Z) \mid g(Z) > \gamma$. Then independently sample Z_i from the distribution of $Z \mid g(Z) = G_i$.

In a moment, we will show that this produces $(Z_1, ..., Z_m)$ from the correct distribution. But first we observe that from this the lemma follows. This is because $\mathbf{Z}_{\text{out}} = (Z_i)_{i \notin \mathcal{I}}$ is realized in Step 2, while $\mathbf{Z}_{\text{in}}$ is precisely $(Z_i)_{i \in \mathcal{I}}$, whose conditional distribution given $(\mathbf{Z}_{\text{out}}, \gamma, \mathcal{I})$ is specified in Step 3. There are m' elements in $(Z_i)_{i \in \mathcal{I}}$, and each is independently obtained by first sampling G from $g(Z) \mid g(Z) > \gamma$, and then sampling Z from $Z \mid g(Z) = G$. But this is the same as sampling Z from $Z \mid g(Z) > \gamma$. This gives us the desired form for the distribution of $\mathbf{Z}_{\text{in}}$ given $(\mathbf{Z}_{\text{out}}, \gamma, \mathcal{I})$, but earlier we showed that $\sigma(\mathbf{Z}_{\text{out}}, \gamma, \mathcal{I}) = \sigma(\mathbf{Z}_{\text{out}}, \mathcal{I})$ and that $\mathbf{Z}_{\text{in}}$ is independent of $\mathcal{I}$ given $\mathbf{Z}_{\text{out}}$. So this gives the same characterization for the distribution of $\mathbf{Z}_{\text{in}}$ given $\mathbf{Z}_{\text{out}}$, and it has the desired form.

Now we show that $(Z_1, ..., Z_m)$ sampled in this way produces samples from the correct distribution. Since each Z_i is independently sampled from the distribution of $Z \mid g(Z) = G_i$, it suffices to show that $(G_1, ..., G_m)$ is an i.i.d. sample from the distribution of $(g(Z_1), ..., g(Z_m))$. To see this, note that under the correct distribution, (i) $g(Z_i) = \gamma$ for $i \notin \mathcal{I}$ surely, and (ii) conditional on $(\gamma, \mathcal{I})$, the correct distribution of the subvector $(g(Z_i))_{i \in \mathcal{I}}$ is that it is an i.i.d. draw of $|\mathcal{I}|$ points from the distribution of $g(Z) \mid g(Z) > \gamma$ by Lemma D.1 (letting $S_i := g(Z_i)$). Then noting that in our sampler $(\gamma, \mathcal{I})$ has the correct distribution by assumption, and that both of the aforementioned properties (i) and (ii) are respected by the way our sampler constructs $(G_1, ..., G_m)$ conditionally on $(\gamma, \mathcal{I})$, we conclude that $(G_1, ..., G_m)$ indeed follows the distribution of $(g(Z_1), ..., g(Z_m))$. $\qquad \square$

Now we may establish Theorem 3.2. In essence, it is simply a re-expression of Lemma D.2, but applied conditionally on $\mathcal{F}$.

D.1.2 Proof of Theorem 3.2

Proof. By assumption, $(X_i, Y_i)_{i:X_i \in \mathcal{R}}$ is an i.i.d. sample from the distribution of $(X, Y) \mid X \in \mathcal{R}$ conditionally on $\mathcal{F}$. Let c and f be as in Algorithm 2 and let $g(x, y) = f(x)$. Note that c and f are fixed conditionally on $\mathcal{F}$. Now, we may apply Lemma D.2 to $(X_i, Y_i)_{i:X_i \in \mathcal{R}}$ conditionally on $\mathcal{F}$. First note that by definition,

$$\{i : X_i \in \mathcal{R} \text{ and } g(X_i, Y_i) \leq \gamma\} = \{i : X_i \in \mathcal{R} \setminus \tilde{\mathcal{R}}\}$$
$$\{i : X_i \in \mathcal{R} \text{ and } g(X_i, Y_i) > \gamma\} = \{i : X_i \in \tilde{\mathcal{R}}\}$$

where we have used the fact that $g(X_i, Y_i) = f(X_i)$. Then in Lemma D.2, $\mathbf{Z}_{\text{out}}$ corresponds to $(X_i, Y_i)_{i:X_i \in \mathcal{R} \setminus \tilde{\mathcal{R}}}$, while $\mathbf{Z}_{\text{in}}$ corresponds to $(X_i, Y_i)_{i:X_i \in \tilde{\mathcal{R}}}$. Lemma D.2, applied conditionally on $\mathcal{F}$, thus allows us to conclude the following:

1. $\gamma = c \wedge \min_{i:X_i \in \mathcal{R}} f(X_i)$ is $\sigma((X_i, Y_i)_{i:X_i \in \mathcal{R} \setminus \tilde{\mathcal{R}}})$-measurable conditionally on $\mathcal{F}$, and is thus $\tilde{\mathcal{F}}$-measurable. Since $\tilde{\mathcal{R}}$ is a function of $\mathcal{R}$, f, c, and γ, all of which are $\tilde{\mathcal{F}}$-measurable, $\tilde{\mathcal{R}}$ is also $\tilde{\mathcal{F}}$-measurable.

2. Similarly, $n(\tilde{\mathcal{R}})$ corresponds to m' in the statement of Lemma D.2, and is thus $\tilde{\mathcal{F}}$-measurable.

3. The distribution of $(X_i, Y_i)_{i:X_i \in \tilde{\mathcal{R}}}$ conditional on $\tilde{\mathcal{F}}$ is $n(\tilde{\mathcal{R}})$ i.i.d. draws from the distribution of $(X, Y) \mid \{X \in \mathcal{R} \text{ and } f(X) > \gamma\}$. But $X \in \mathcal{R}$ and $f(X) > \gamma$ if and only if $X \in \tilde{\mathcal{R}}$, so this is equivalent to the distribution of $(X, Y) \mid X \in \tilde{\mathcal{R}}$, as desired.

$\qquad \square$

D.1.3 Proof of Corollary 3.3

Proof. The first part of the following corollary follows directly from recursively applying Theorem 3.2. For the second part, write $Z := (X_i, Y_i)_{i:X_i \in \mathcal{R}_\nu}$, and $Z_t := (X_i, Y_i)_{i:X_i \in \mathcal{R}_t}$. Let

$h : \mathbb{R}^n \to \mathbb{R}$ be bounded and measurable. We will allow $h(\cdot)$ to be applied to a vector z of length less than n by simply padding z with an appropriate number of zeros. Then

$$\mathbb{E}[h(Z) \mid \mathcal{F}_\nu] = \sum_{t=0}^{\infty} \mathbb{E}[h(Z) \mid \mathcal{F}_t] \mathbb{1}\{\nu = t\}$$

$$= \sum_{t=0}^{\infty} \mathbb{E}[h(Z) \mathbb{1}\{\nu = t\} \mid \mathcal{F}_t]$$

$$= \sum_{t=0}^{\infty} \mathbb{E}[h(Z_t) \mathbb{1}\{\nu = t\} \mid \mathcal{F}_t]$$

$$= \sum_{t=0}^{\infty} \mathbb{E}[h(Z_t) \mid \mathcal{F}_t] \mathbb{1}\{\nu = t\}.$$

The first equality follows from Lemma F.4, and the remainder follows from noting that $\mathbb{1}\{\nu = t\}$ is $\mathcal{F}_t$-measurable (as it is a stopping time) and that $h(Z)\mathbb{1}\{\nu = t\} = h(Z_t)\mathbb{1}\{\nu = t\}$. Thus, for almost all $\omega \in \Omega$,

$$\mathbb{E}[h(Z) \mid \mathcal{F}_\nu](\omega) = \sum_{t=0}^{\infty} \mathbb{E}[h(Z_t) \mid \mathcal{F}_t](\omega) \mathbb{1}\{\nu(\omega) = t\} = \mathbb{E}[h(Z_{\nu(\omega)}) \mid \mathcal{F}_{\nu(\omega)}](\omega).$$

The second equality follows from noting that the sum is nonzero only when $t = \nu(\omega)$. Since conditional on $\mathcal{F}_{\nu(\omega)}$, $Z_{\nu(\omega)}$ is almost surely of length $m(\omega) := \{|i : X_i(\omega) \in \mathcal{R}_{\nu(\omega)}|\}$, for any fixed ω it suffices to consider functions $h : \mathbb{R}^{m(\omega)} \to \mathbb{R}$, where now we drop the zero padding convention. In particular, the above holds for $h(z) := \mathbb{1}\{z \in B\}$ for measurable $B \subseteq \mathbb{R}^{m(\omega)}$. But this shows that the distribution of Z given $\mathcal{F}_\nu$ at ω is the same as the distribution of Z_t given $\mathcal{F}_t$ at ω when we let $t = \nu(\omega)$. The latter we have already characterized as having the distribution of $|\{i : X_i(\omega) \in \mathcal{R}_t(\omega)\}|$ i.i.d. draws from $(X, Y) \mid X \in \mathcal{R}_t(\omega)$. Simply plugging in $t = \nu(\omega)$ yields the conclusion. □

D.2 Error control for interactive testing

We prove a generalization of Lemma 3.7 which allows for slight discrepancies between the achieved level and the target levels of the conditional tests. This will be useful when we turn to asymptotic arguments later, and Lemma 3.7 follows as a special case.

Lemma D.3 (Approximate sequential error control). *Let* $(\mathcal{H}_t, \tilde{\alpha}_t, \phi_t)_{t=0}^m$ *be a conditionally valid testing sequence with respect to a filtration* $(\mathcal{G}_t)_{t=0}^m$, *and suppose* $\tilde{\alpha}_t = \alpha_t + \delta_t$ *where* $\alpha_t \in [0, 1]$ *and* δ_t *are also* $\mathcal{G}_t$-*measurable. We will think of* α_t *as the target levels and* $\tilde{\alpha}_t$ *as the achieved levels. Define the first rejected index* $\tau := \min\{t : \phi_t = 1\}$ *with* $\tau = \infty$ *if the set is empty, and let* $\mathcal{H}_\infty = 1$ *by convention. If the target levels satisfy* $1 - \prod_{t=0}^m (1 - \alpha_t) \leq \alpha$ *almost surely, then rejecting* $\mathcal{H}_\tau$ *approximately controls the Type I error rate:*

$$\mathbb{P}(\mathcal{H}_\tau = 0) \leq \alpha + \mathbb{E}[\epsilon_0] \tag{13}$$

where ϵ_0 *is a linear combination of* $(\delta_t)_{t=0}^m$ *with random coefficients bounded between* 0 *and* 1. *Furthermore, if* $(\mathcal{H}_t, \tilde{\alpha}_t, \phi_t)_{t=0}^m$ *is conditionally exact,* $1 - \prod_{t=0}^m (1 - \alpha_t) = \alpha$, *and* $\mathcal{H}_t = 1$

69

almost surely for all $t = 0, ..., m$ (i.e. we are under the global null), then Equation (13) *holds with equality.*

Note that Lemma 3.7 is an immediate corollary of Lemma D.3 by letting $\delta_t = 0$ for all t, whence $\mathbb{E}[\epsilon_0] = 0$. Also note the following asymptotic corollary which places explicit growth conditions on the approximation errors $(\delta_t)_{t=0}^m$. To state the asymptotics, we need to define a sequence of conditionally valid testing sequences indexed by n. That is, let $(\mathcal{H}_t^{(n)}, \tilde{\alpha}_t^{(n)}, \phi_t^{(n)})_{t=0}^{m(n)}$ be a conditionally valid testing sequence with respect to $(\mathcal{G}_t^{(n)})_{t=0}^{m(n)}$ for each $n \geq 1$, and note that the total number of tests $m(n)$ can vary with n. We will also write $\tilde{\alpha}_t^{(n)} = \alpha_t^{(n)} + \delta_t^{(n)}$. Let $\tau^{(n)}$ be the first rejected index for the nth conditionally valid testing sequence, and by slight abuse of notation write $\mathcal{H}_\tau^{(n)} = \mathcal{H}_{\tau^{(n)}}^{(n)}$.

Corollary D.4 (Asymptotic sequential error control). *Suppose that* $(\mathcal{H}_t^{(n)}, \tilde{\alpha}_t^{(n)}, \phi_t^{(n)})_{t=0}^{m(n)}$, $(\mathcal{G}_t^{(n)})_{t=0}^{m(n)}$, *and* $(\alpha_t^{(n)}, \delta_t^{(n)})_{t=0}^{m(n)}$ *satisfy the conditions of Lemma D.3 for each n. Furthermore, suppose that* $|\delta_0^{(n)}| + ... + |\delta_{m(n)}^{(n)}| \xrightarrow{L^1} 0$ *as $n \to \infty$. Then*

$$\limsup_{n \to \infty} \mathbb{P}(\mathcal{H}_\tau^{(n)} = 0) \leq \alpha.$$

In the case where we satisfy the exactness conditions in Lemma D.3, we have

$$\lim_{n \to \infty} \mathbb{P}(\mathcal{H}_\tau^{(n)} = 0) = \alpha.$$

Corollary D.4 follows directly from Lemma D.3 by noting that since ϵ_0 is a linear combination of $(\delta_t)_{t=0}^m$ with random coefficients bounded between 0 and 1, then (introducing an n superscript on ϵ_0) we have the bound

$$|\mathbb{E}[\epsilon_0^{(n)}]| \leq \mathbb{E}[|\epsilon_0^{(n)}|] \leq \mathbb{E}[|\delta_0^{(n)}| + ... + |\delta_{m(n)}^{(n)}|].$$

Both the limit and the limit superior of the right-hand side go to 0 as $n \to \infty$ by assumption, and the result follows.

Finally, we prove the centerpiece of this section.

Proof of Lemma D.3. First note that we may suppose without loss of generality that $1 - \prod_{t=0}^m (1 - \alpha_t) = \alpha$. If not, we can simply define an additional hypothesis $\mathcal{H}_{m+1} = 0$, test $\phi_{m+1} = 0$, and target level $\alpha_{m+1} = 1 - (1 - \alpha) \prod_{t=0}^m (1 - \alpha_t)^{-1}$ with no approximation error, i.e. $\delta_{m+1} = 0$ and $\tilde{\alpha}_{m+1} = \alpha_{m+1}$. Then since this additional test accepts almost surely, including it will not change the value of $\mathbb{P}(\mathcal{H}_\tau = 0)$. Thus, we may apply the proof below to $(\mathcal{H}_t, \tilde{\alpha}_t, \phi_t)_{t=0}^{m+1}$, which is still a conditionally valid testing sequence, and is conditionally exact if $(\mathcal{H}_t, \tilde{\alpha}_t, \phi_t)_{t=0}^m$ was conditionally exact (because then $\alpha_{m+1} = 0$ almost surely).

Define $A_t := 1 - \phi_t(1 - \mathbb{1}\{\tau < t\})$ and $S_t := 1 - (1 - \mathcal{H}_t)(1 - A_t)$. A_t is 1 if the tth hypothesis is not reported (A for "accept") either because ϕ_t did not reject or because there was an earlier rejection, and S_t represents no mistake at stage t (S for "safe"), which happens if either we don't report the tth hypothesis as a rejection or if the tth hypothesis is not null (i.e. $\mathcal{H}_t = 1$, we are under the alternative). Also define $\tilde{\mathcal{G}}_t := \sigma(\mathcal{G}_t, (\phi_s)_{s<t})$. Note that since $\tau < t$ if and only if $\max_{s<t} \phi_s = 1$, then A_t, S_t are $\tilde{\mathcal{G}}_{t+1}$-measurable.

First we will show that $\mathbb{E}[S_t \mid \tilde{\mathcal{G}}_t] \geq 1 - \alpha_t - \delta_t$ almost surely for all t. We can see this by cases.

1. **Case 1:** Condition on $\{A_0 = ... = A_{t-1} = 0\} \cap \{\mathcal{H}_t = 0\} \in \tilde{\mathcal{G}}_t$, i.e. that there has not been a rejection yet and the null hypothesis is true. Conditional on this, $S_t = A_t = 1 - \phi_t$ by construction, and hence $\mathbb{E}[S_t \mid \tilde{\mathcal{G}}_t] \geq 1 - \tilde{\alpha}_t = 1 - \alpha_t - \delta_t$ by the definition of conditionally valid testing sequences (note that here we are conditioning on both $\mathcal{H}_t = 0$ and $\max_{s<t} \phi_s = 0$ which is implied by $A_0 = ... = A_{t-1} = 0$).

2. **Case 2:** Condition on $\{A_s = 1 \text{ for some } s < t\} \cup \{\mathcal{H}_t = 1\} \in \tilde{\mathcal{G}}_t$, i.e. that there has been a rejection or the null hypothesis is false. Conditional on this, we either have that $\tau < t$ and hence $A_t = 1$, or $\mathcal{H}_t = 1$. Either way, we see that $S_t = 1$ almost surely, and in particular $\mathbb{E}[S_t \mid \tilde{\mathcal{G}}_t] = 1 \geq 1 - \tilde{\alpha}_t = 1 - \alpha_t - \delta_t$ since $\tilde{\alpha}_t \in [0,1]$.

Since these two cases are exhaustive, we conclude that $\mathbb{E}[S_t \mid \tilde{\mathcal{G}}_t] \geq 1 - \alpha_t - \delta_t$ almost surely for all t.

Now, define $\epsilon_m := \delta_m$ and recursively define for $t = m-1, ..., 0$,

$$\epsilon_t := \delta_t (1 - \alpha) \prod_{s=0}^{t} (1 - \alpha_s)^{-1} + S_t \epsilon_{t+1}.$$

Note that ϵ_t is a linear combination of $\delta_t, ..., \delta_m$ with random coefficients bounded between 0 and 1. Also define $\epsilon_{m+1} = 0$ as a convention. We will show that

$$\mathbb{E}\left[S_0 ... S_t \left((1 - \alpha) \prod_{s=0}^{t} (1 - \alpha_s)^{-1} - \epsilon_{t+1}\right)\right] \geq \mathbb{E}\left[S_0 ... S_{t-1} \left((1 - \alpha) \prod_{s=0}^{t-1} (1 - \alpha_s)^{-1} - \epsilon_t\right)\right] \tag{14}$$

for all $t = 0, ..., m$, where we use the convention that empty products evaluate to 1. To see

this, note that

$$
\mathbb{E}\left[S_0...S_t\left((1-\alpha)\prod_{s=0}^{t}(1-\alpha_s)^{-1} - \epsilon_{t+1}\right)\right]
$$

$$
=\mathbb{E}\left[S_0...S_{t-1}\mathbb{E}\left[\left(S_t(1-\alpha)\prod_{s=0}^{t}(1-\alpha_s)^{-1} - S_t\epsilon_{t+1}\right)\,\middle|\,\tilde{\mathcal{G}}_t\right]\right]
$$

$$
=\mathbb{E}\left[S_0...S_{t-1}\left(\mathbb{E}\left[S_t\,\middle|\,\tilde{\mathcal{G}}_t\right](1-\alpha)\prod_{s=0}^{t}(1-\alpha_s)^{-1} - \mathbb{E}\left[S_t\epsilon_{t+1}\,\middle|\,\tilde{\mathcal{G}}_t\right]\right)\right]
$$

$$
\geq\mathbb{E}\left[S_0...S_{t-1}\left((1-\alpha_t-\delta_t)(1-\alpha)\prod_{s=0}^{t}(1-\alpha_s)^{-1} - \mathbb{E}\left[S_t\epsilon_{t+1}\,\middle|\,\tilde{\mathcal{G}}_t\right]\right)\right]
$$

$$
=\mathbb{E}\left[S_0...S_{t-1}\mathbb{E}\left[(1-\alpha_t-\delta_t)(1-\alpha)\prod_{s=0}^{t}(1-\alpha_s)^{-1} - S_t\epsilon_{t+1}\,\middle|\,\tilde{\mathcal{G}}_t\right]\right]
$$

$$
=\mathbb{E}\left[S_0...S_{t-1}\mathbb{E}\left[(1-\alpha)\prod_{s=0}^{t-1}(1-\alpha_s)^{-1} - \delta_t(1-\alpha)\prod_{s=0}^{t}(1-\alpha_s)^{-1} - S_t\epsilon_{t+1}\,\middle|\,\tilde{\mathcal{G}}_t\right]\right]
$$

$$
=\mathbb{E}\left[S_0...S_{t-1}\mathbb{E}\left[(1-\alpha)\prod_{s=0}^{t-1}(1-\alpha_s)^{-1} - \epsilon_t\,\middle|\,\tilde{\mathcal{G}}_t\right]\right]
$$

$$
=\mathbb{E}\left[S_0...S_{t-1}\left((1-\alpha)\prod_{s=0}^{t-1}(1-\alpha_s)^{-1} - \epsilon_t\right)\right].
$$

(15)

To go to the second line of the above, we use the fact that $S_0,...,S_{t-1}$ are $\tilde{\mathcal{G}}_t$-measurable and hence can be moved out of the inner conditional expectation. The third line follows from linearity of expectation and the fact that α_s is $\tilde{\mathcal{G}}_t$-measurable for $s \leq t$. The fourth line follows from $\mathbb{E}[S_t \mid \tilde{\mathcal{G}}_t] \geq 1 - \alpha_t - \delta_t$ as we showed above. The fifth line follows from moving the central term back inside of the conditional expectation, which we can do since it consists of $\tilde{\mathcal{G}}_t$-measurable quantities. The sixth line follows from algebraic manipulations and the seventh line follows from the definition of ϵ_t. The last line follows from moving the $S_0...S_{t-1}$ back inside of the conditional expectation and applying the law of total expectation. This establishes Equation (14).

Finally, note that the right-hand side of Equation (14) is simply the left-hand side but with $t-1$ substituted for t. Thus, by forming the chain of inequalities implied by Equation (14) beginning with $t = m$ and ending with $t = 0$, we obtain

$$
\mathbb{E}[S_0...S_m] \geq \mathbb{E}[(1-\alpha) - \epsilon_0] = 1 - \alpha - \mathbb{E}[\epsilon_0].
$$

Since $\mathbb{P}(\mathcal{H}_\tau = 0) = 1 - \mathbb{E}[S_0...S_m]$, this proves the first part of the lemma.

Now we turn our attention to the second part of the lemma. Recall that we have already assumed without loss of generality that $1 - \prod_{t=0}^{m}(1-\alpha_t) = \alpha$. We additionally assume that the testing sequence is conditionally exact and $\mathcal{H}_t = 0$ a.s. for $t = 0,...,m$. We argue that under these additional assumptions, the inequality in Equation (14) can be turned into an equality. In particular, we argue that the sole inequality in Equation (15) can be turned into

an equality. The reason is that now $S_0...S_{t-1}\mathbb{E}[S_t \mid \tilde{\mathcal{G}}_t] = S_0...S_{t-1}(1 - \alpha_t - \delta_t)$ almost surely. We again proceed by cases. Conditional on $A_s = 0$ for some $s < t$, both sides of the expression are 0 since $\mathcal{H}_s = 0$ for all s and thus $S_s = 0$ for some $s < t$. Conditional on the complement, $S_0...S_{t-1} = 1$ and the expression reduces to $\mathbb{E}[S_t \mid \tilde{\mathcal{G}}_t] = 1 - \alpha_t - \delta_t$. But this holds by the definition of conditional exactness. Using $S_0...S_{t-1}\mathbb{E}[S_t \mid \tilde{\mathcal{G}}_t] = S_0...S_{t-1}(1 - \alpha_t - \delta_t)$ lets us manipulate the third line into the fourth line of Equation (15) with equality. Hence, Equation (14) holds with equality, and the remainder of the proof is the same with "$\geq$" signs replaced with "$=$" signs. $\qquad\square$

D.3 Validity results for binary outcomes

D.3.1 Proof of Theorem 3.9

Proof. We will make use of Lemma 3.7. Let $(\mathcal{R}_t, \alpha_t, \mathcal{F}_t)_{t=0}^m$ be defined as in Algorithm 1. Let $\mathcal{H}_t := \mathbb{1}\{\mu(\mathcal{R}_t) > \mu_{\text{cut}}\}$, and let ϕ_t and C_t be defined as in the theorem statement. Let $\mathcal{G}_t := \sigma(\mathcal{F}_t, (U_s)_{s<t})$ where U_t is the auxiliary random variable (for instance, a standard uniform independent of everything else) used to sample from $Q(1 - \alpha_t; n_t, \mu_{\text{cut}}, n_t M_t)$ in the definition of C_t; in particular, C_t is $\sigma(\mathcal{F}_t, (U_s)_{s\leq t})$-measurable and M_t is $\mathcal{G}_t$-measurable. Then by Lemma 3.7 it suffices to show that $(\mathcal{H}_t, \alpha_t, \phi_t)_{t=0}^m$ is conditionally valid, and conditionally exact under the additional assumptions (see Definition 3.6).

Let t be fixed. We need to show that

$$\mathbb{P}(\bar{\mu}_t \leq C_t \mid \sigma(\mathcal{G}_t, (\phi_s)_{s<t})) \geq 1 - \alpha_t \quad \text{conditional on} \quad \{\mathcal{H}_t = 0 \text{ and } \max_{s<t} \phi_s = 0\}.$$

Implicitly condition on $\{\mathcal{H}_t = 0 \text{ and } \max_{s<t} \phi_s = 0\}$ throughout. We rewrite the left-hand side,

$$
\begin{aligned}
\mathbb{P}(\bar{\mu}_t \leq C_t \mid \mathcal{G}_t, (\phi_s)_{s<t}) &= \mathbb{P}(\bar{\mu}_t \leq C_t \mid \mathcal{G}_t, \max_{s<t} \phi_s) \\
&= \mathbb{P}(\bar{\mu}_t \leq C_t \mid \mathcal{G}_t, \max_{s<t} \phi_s = 0) \\
&= \frac{\mathbb{P}(\bar{\mu}_t \leq C_t, \max_{s<t} \phi_s = 0 \mid \mathcal{G}_t)}{\mathbb{P}(\max_{s<t} \phi_s = 0 \mid \mathcal{G}_t)} \\
&= \frac{\mathbb{P}(\bar{\mu}_t \leq C_t, \bar{\mu}_t \leq M_t \mid \mathcal{G}_t)}{\mathbb{P}(\bar{\mu}_t \leq M_t \mid \mathcal{G}_t)} \\
&= \frac{\mathbb{P}(\bar{\mu}_t \leq C_t \mid \mathcal{F}_t, (U_s)_{s<t})}{\mathbb{P}(\bar{\mu}_t \leq M_t \mid \mathcal{F}_t, (U_s)_{s<t})}.
\end{aligned}
$$

The first equality follows from the fact that $\max_{s<t} \phi_s = 0 \iff \phi_1 = ... = \phi_{t-1} = 0$, so we may apply Lemma F.3. The second equality follows from Lemma F.2, and the third equality is by definition. The fourth equality follows from the fact that $\max_{s<t} \phi_s = 0$ if and only if $\bar{\mu}_t \leq M_t$ (recalling the calculations in Appendix C.7). The last equality follows from expanding $\mathcal{G}_t$ and the fact that $C_t \leq M_t$ almost surely by definition. Now consider a fixed $\omega \in \{\mathcal{H}_t = 0 \text{ and } \max_{s<t} \phi_s = 0\}$. Then

$$\mathbb{P}(\bar{\mu}_t \leq C_t \mid \mathcal{F}_t, (U_s)_{s<t})(\omega) = \mathbb{P}(n_t \bar{\mu}_t \leq n_t C_t \mid \mathcal{F}_t, (U_s)_{s<t})(\omega) = \mathbb{P}(Z \leq C') \quad (16)$$

where $Z \sim \text{Binom}(n_t(\omega), \mu(\mathcal{R}_t(\omega)))$ and C' is the independently randomized output of $Q(1 - \alpha_t(\omega); n_t(\omega), \mu(\mathcal{R}_t(\omega)), n_t(\omega) M_t(\omega))$. This follows from noticing a few things. First,

$\alpha_t, n_t, \mathcal{R}_t$, and M_t are $\sigma(\mathcal{F}_t, (U_s)_{s<t})$-measurable and hence we can refer to their fixed values conditionally on $\sigma(\mathcal{F}_t, (U_s)_{s<t})$ at ω unambiguously; this is what the notation $\alpha_t(\omega)$ means, and so on. Next, $n_t\bar{\mu}_t$ is the sum of the Y values of points that fall in $\mathcal{R}_t(\omega)$, and Corollary 3.3 tells us that conditionally on $\sigma(\mathcal{F}_t, (U_s)_{s<t})$, each of these points is i.i.d. Bernoulli with rate $\mu(\mathcal{R}_t(\omega))$ (note that the U_s are independent of everything and thus do not affect the application of Corollary 3.3). Thus, $n_t\bar{\mu}_t$ is distributed as $\text{Binom}(n_t(\omega), \mu(\mathcal{R}_t(\omega)))$ conditionally on $\sigma(\mathcal{F}_t, (U_s)_{s<t})$ at ω. Next, $n_t C_t$ is by definition the independently randomized output of $Q(1 - \alpha_t(\omega); n_t(\omega), \mu(\mathcal{R}_t(\omega)), n_t(\omega)M_t(\omega))$ given $\sigma(\mathcal{F}_t, (U_s)_{s<t})$ at ω, and in particular it is conditionally independent of $n_t\bar{\mu}_t$. With these observations, we see that the right-hand side of Equation (16) is merely a re-expression of the middle term.

Following the same logic, we obtain that

$$\mathbb{P}(\bar{\mu}_t \le M_t \mid \mathcal{F}_t, (U_s)_{s<t})(\omega) = \mathbb{P}(Z \le M')$$

where $M' = M_t(\omega)$. Putting this all together, we obtain that

$$\mathbb{P}(\bar{\mu}_t \le C_t \mid \mathcal{G}_t)(\omega) = \frac{\mathbb{P}(Z \le C')}{\mathbb{P}(Z \le M')} = \mathbb{P}(Z \le C' \mid Z \le M')$$

for almost all $\{\mathcal{H}_t = 0 \text{ and } \max_{s<t} \phi_s = 0\}$, where we have again used the fact that $C' \le M'$ a.s. Recalling that $\mathcal{H}_t(\omega) = 0 \iff \mu(\mathcal{R}_t(\omega)) \le \mu_{\text{cut}}$, then by stochastic domination for truncated binomials (Lemma F.6) we have that the above continues simplifying as

$$\cdots = \mathbb{P}(Z \le C' \mid Z \le M') \ge \mathbb{P}(Z' \le C' \mid Z' \le M') \tag{17}$$

where $Z' \sim \text{Binom}(n_t(\omega), \mu_{\text{cut}})$. But $\mathbb{P}(Z' \le C' \mid Z' \le M') = 1 - \alpha_t(\omega)$ by definition, and thus we conclude that $\mathbb{P}(\bar{\mu}_t \le C_t \mid \mathcal{G}_t) \ge 1 - \alpha_t$ a.s. conditionally on $\{\mathcal{H}_t = 0 \text{ and } \max_{s<t} \phi_s = 0\}$ as desired. This shows conditional validity, and as mentioned an application of the first part of Lemma 3.7 completes the first part of the theorem.

To show conditional exactness under the additional assumptions, we note that $\mu(X) = \mu_{\text{cut}}$ a.s. implies that $\mu(\mathcal{R}_t(\omega)) = \mu_{\text{cut}}$ almost surely. Thus, the inequality in Equation (17) can be replaced with an equality, allowing us to conclude that $\mathbb{P}(\bar{\mu}_t \le C_t \mid \mathcal{G}_t) = 1 - \alpha_t$ a.s. conditional on $\{\mathcal{H}_t = 0 \text{ and } \max_{s<t} \phi_s = 0\}$. Then we have conditional exactness, and an application of the second part of Lemma 3.7 establishes the second part of the theorem. $\square$

D.4 Validity results for general outcomes

In Section 3.6, we stated a simpler version of our asymptotic results for expositional clarity. Here, we will prove a more general set of claims. We briefly motivate the generalization. First of all, we relax the asymptotic conditions that are stated in the main text. Moreover, this relaxation allows us to state uniform guarantees (or equivalently, guarantees over a triangular array where the base distribution may change from row to row). Finally, we relax the requirement that the minimum sample size allocated to any non-trivial test must be a constant proportion p_{prop} of n. The reason this last relaxation is appealing is that the original condition essentially precludes the discovery of a region $\mathcal{R}$ whose probability mass $\rho(\mathcal{R}) := \mathbb{P}(X \in \mathcal{R})$ is much less than p_{prop}, since $\rho(\mathcal{R}_t) \approx n_t/n$. This is generally not of major concern in key application areas such as RCTs, since the study populations are typically

designed with a strong prior belief of efficacy for most if not all units in the population. Moreover, RCT sample sizes are typically modest so that extremely small subgroups will typically lack sufficient sample sizes to justify approximate inference via CLT. Nonetheless, there may be situations where one hopes to identify an extremely small subgroup from a very large dataset, perhaps arising in observational causal inference or regression settings.

We generalize both the construction of the tests and the asymptotic conditions; these are given in Appendices D.4.1 and D.4.2 respectively. Our generalized claims are stated at the end of the latter, and we devote Appendices D.4.3–D.4.6 to proving them. We end with a section describing primitive conditions—including those given in Section 3.6—which imply our generalized conditions (Appendix D.4.7), and a somewhat stylized discussion of why we expect our generalization to be powerful in practice (Appendix D.4.8).

Finally, we introduce some notation that we will use throughout this section. Recall that for a fixed $\mathcal{R} \subseteq \mathcal{X}$ we have defined $\rho(\mathcal{R}) = \mathbb{P}(X \in \mathcal{R})$. For any fixed $\mathcal{R} \subseteq \mathcal{X}$ and positive integer k, recall that $\mu(\mathcal{R}) = \mathbb{E}[Y \mid X \in \mathcal{R}]$ and define

$$
\begin{aligned}
\rho(\mathcal{R}) &:= \mathbb{P}(X \in \mathcal{R}), \\
\bar{\mu}_{k,\mathrm{raw}}(\mathcal{R}) &:= \mathbb{E}[|Y|^k \mid X \in \mathcal{R}], \\
\bar{\mu}_{k,\mathrm{cent}}(\mathcal{R}) &:= \mathbb{E}[|Y - \mu(\mathcal{R})|^k \mid X \in \mathcal{R}], \\
\sigma_k(\mathcal{R}) &:= \sqrt{\mathrm{Var}(Y^k \mid X \in \mathcal{R})}.
\end{aligned}
\tag{18}
$$

We will also write $\sigma(\mathcal{R}) = \sigma_1(\mathcal{R})$ for convenience.

D.4.1   Formalizing the oracle and empirical tests

The proof strategy we will take is to describe an *oracle* test which standardizes the critical values using the true variance of the mean estimate. Because the true variance of $\bar{\mu}_t$ is a property of the region $\mathcal{R}_t$, we have that this true variance is $\mathcal{F}_t$-measurable and thus easier to analyze. We then show that the test which standardizes via the empirical variance produces the same output as this oracle test with probability tending to 1 asymptotically. We devote this section to setting up the proof framework.

Let $(\mathcal{R}_t, \alpha_t, \mathcal{F}_t)_{t=0}^m$ be as defined in Algorithm 1. For a fixed region $\mathcal{R} \subseteq \mathcal{X}$ define $\sigma^2(\mathcal{R}) := \mathrm{Var}(Y \mid X \in \mathcal{R})$, and define

$$
V_t := \sigma^2(\mathcal{R}_t).
$$

We will now use the variables M_t and C_t slightly differently from how they are presented in Definition 3.10, but which are more convenient to work with. First, define the following *oracle intermediate statistics* for $s \leq t$:

$$
\Delta_{s,t} := \frac{1}{\sqrt{V_s} \cdot \sqrt{n_s}} \sum_{i:X_i \in \mathcal{R}_s \setminus \mathcal{R}_t} Y_i \quad \text{and} \quad v_{s,t} := \sqrt{\frac{V_t}{V_s}} \sqrt{\frac{n_t}{n_s}}.
\tag{19}
$$

Define the *empirical intermediate statistics* $\hat{\Delta}_{s,t}$ and $\hat{v}_{s,t}$ identically to the above, but with V_s replaced with $\hat{V}_s$ and V_t replaced with $\hat{V}_t$. These quantities track various statistics that will be used to relate the test statistic at stage s to the test statistic at stage t. We use

the notation $\Phi^{-1}(q; M) := \Phi^{-1}(q \cdot \Phi(M))$ to denote the qth quantile of $\mathcal{N}(0,1)$ truncated to be no greater than M. We will also define $\Phi(x; M) = \Phi(x)/\Phi(M)$ for $x \leq M$ and $\Phi(x; M) = 1$ otherwise. This is the CDF of a $\mathcal{N}(0,1)$ truncated to be no greater than M, so that $\Phi(\Phi^{-1}(q; M); M) = q$. We define two versions of the test statistic and critical values: an *oracle* version which has access to the random variables V_t and an *empirical* version that uses $\hat{V}_t$ in place of V_t wherever the latter appears. We also describe these tests in greater generality than what is described in the main text and will elaborate on the connection momentarily. Throughout, let $p_{\text{prop}} \in (0,1)$ be fixed.

Definition D.5 (Oracle normal test statistic and critical values). *Define $T_t := \sqrt{n_t} \cdot V_t^{-1/2} \bar{\mu}_t$. Also, recursively define*

$$M_t := \min_{s < t : n_t/n_s \geq p_{\text{prop}}} \left\{ \frac{C_s - \Delta_{s,t}}{v_{s,t}} \right\} \quad and \quad C_t := \max\{0, \Phi^{-1}(1 - \alpha_t; M_t)\}.$$

where the minimum is ∞ if the set is empty (i.e. when $t = 0$). The oracle tests are defined as $\phi_t := \mathbb{1}\{T_t > C_t\}$.

Definition D.6. *(Empirical normal test statistics and critical values) Define $\hat{T}_t := \sqrt{n_t} \cdot \hat{V}_t^{-1/2} \bar{\mu}_t$. Also, recursively define*

$$\hat{M}_t := \min_{s < t : n_t/n_s \geq p_{\text{prop}}} \left\{ \frac{\hat{C}_s - \hat{\Delta}_{s,t}}{\hat{v}_{s,t}} \right\} \quad and \quad \hat{C}_t := \max\{0, \Phi^{-1}(1 - \alpha_t; \hat{M}_t)\}$$

where the minimum is ∞ if the set is empty (i.e. when $t = 0$). The empirical tests are defined as $\hat{\phi}_t := \mathbb{1}\{T_t > \hat{C}_t\}$.

Note that by construction, $\Phi(C_t; M_t) \geq 1 - \alpha_t$, and the same is true when (C_t, M_t) is replaced with $(\hat{C}_t, \hat{M}_t)$. Also, $V_t, \Delta_{s,t}, v_{s,t}, M_t$, and C_t are $\mathcal{F}_t$-measurable for all $s \leq t$ (though the same is not true of the hatted quantities). This important property will allow us to more easily establish the Type I error control for the oracle test.

Now we connect this to what is described in the main text. Suppose Constraint 3.12 were enforced with the same p_{prop} as defined in this section. Then under Constraint 3.12, it is the case that if $\alpha_t > 0$, then $n_t/n_s \geq p_{\text{prop}}$ for all $s < t$. In this case, the behavior of Algorithm 1 is not affected by dropping the condition $n_t/n_s \geq p_{\text{prop}}$ from the definition of M_t in Definition D.5 and $\hat{M}_t$ in Definition D.6. From here, the rationale for these definitions is the same as in Appendix C.7. Since

$$T_t = \frac{T_s - \Delta_{s,t}}{v_{s,t}}$$

for $s \leq t$, then

$$T_s \leq C_s \text{ for all } s < t \iff T_t = \frac{T_s - \Delta_{s,t}}{v_{s,t}} \leq \frac{C_s - \Delta_{s,t}}{v_{s,t}} \text{ for all } s < t$$

$$\iff T_t \leq \min_{s < t} \left\{ \frac{C_s - \Delta_{s,t}}{v_{s,t}} \right\} = M_t.$$

(20)

76

The same is true when (C_t, M_t) is replaced with $(\hat{C}_t, \hat{M}_t)$. In fact, a bit of straightforward algebraic manipulation shows that the empirical test we have described here is equivalent to the test we described in the main body as long as Constraint 3.12 is enforced, in the sense that ϕ_t (based on Definition 3.10 and used in Theorem 3.13) and $\hat{\phi}_t$ (from Definition D.6) are equal almost surely. Thus, the validity guarantees we will give for these tests subsume the validity guarantees stated in the main text.

When Constraint 3.12 is not enforced, the rationale for these definitions is similar. As a matter of fact, Equation (20) still holds up to the last line, but where now instead of having the final quantity equal to M_t we have

$$\min_{s<t}\left\{\frac{C_s - \Delta_{s,t}}{v_{s,t}}\right\} \leq \min_{s<t:n_t/n_s \geq p_{\text{prop}}}\left\{\frac{C_s - \Delta_{s,t}}{v_{s,t}}\right\} = M_t.$$

That is, we truncate more conservatively, which results in a more conservative critical value C_t and, consequently, a more conservative test, but which does not affect validity. This generalization was designed to overcome a technical difficulty in our proofs; namely, that it is somewhat difficult to characterize the effect of truncation for tests that utilize vastly different sample sizes.[9] Thus, these truncation levels are designed to utilize only truncation information from tests whose sample sizes are of the same order as the current sample size. This will allow us to weaken the requirement in Constraint 3.12 that the minimum sample size is a constant proportion of n, allowing chiseling to validly discover regions of arbitrarily small probability mass. In Appendix D.4.8, we give some heuristics to explain why we do not expect to lose substantial power from discarding distal truncation information. Lastly, we note that in some cases this more general construction does not discard distal truncation information at all—for instance, when Constraint 3.12 is satisfied as we have seen, or when we enforce a similar constraint that the non-trivial tests (i.e. where $\alpha_t > 0$) are not spaced too far apart.

Remark D.7 (Undefined edge cases). *As in the main text, if ever a quantity is undefined, any test that depends upon it will be assumed to* accept. *Under assumptions that we will soon state, for t where $\alpha_t > 0$ this edge case will happen with probability going to 0, and hence for simplicity of exposition we will simply assume that all of the above quantities are well-defined almost surely. This will already be the case if the distribution of Y has no discrete atoms and n is sufficiently large, and it is straightforward (though burdensome) to amend our proofs to handle undefined quantities. By, in essence, "intersecting" every event that appears in any probability statement with the event*

$$E := \{\text{all quantities are well-defined at stages } t \text{ such that } \alpha_t > 0\},$$

all of our arguments may be carried out identically, and the same results follow upon noticing that $\mathbb{P}(E) \to 1$ asymptotically.

---

[9]In fact, this is only a difficulty for the empirical tests, which are what we ultimately care about. It is interesting to note that we can prove validity for a version of the oracle test that uses all the truncation information, suggesting that the primary difficulty is in controlling the error of the variance estimation.

D.4.2 Formalizing the asymptotic conditions

We will prove a more general version of Theorem 3.13 under more general conditions. We first state a high-level set of conditions that will suffice to imply validity. These conditions are somewhat abstract, so in Appendix D.4.7 we will describe primitive conditions (including the primitives stated in Section 3.6) that suffice to imply these high-level conditions.

Consider a triangular array of data points $(X_{i,n}, Y_{i,n})_{i=1}^n$ for $n = 1, 2, ...$ where the nth row consists of n i.i.d. data points drawn from the distribution of $(X^{(n)}, Y^{(n)})$. We allow the distribution of the pair of random variables to vary with n. Though $Y^{(n)}$ will always be real-valued, we suppose $X^{(n)} \in \mathcal{X}^{(n)}$ where $\mathcal{X}^{(n)}$ may vary with n. We imagine that chiseling is applied separately to each row of the triangular array. As the number of steps m may vary with n, we write $m(n)$. For each n, this produces chiseled regions $(\mathcal{R}_t^{(n)})_{t=0}^{m(n)}$, levels $(\alpha_t^{(n)})_{t=0}^{m(n)}$, and a selected region which we denote $\mathcal{R}_\tau^{(n)}$. Subsequently, wherever appropriate, quantities will be understood to depend on n unless stated otherwise (for instance, the critical values C_t now also depend on n and should be implicitly understood as $C_t^{(n)}$). For readability, we will often suppress dependence on n in the proofs where it is not confusing, only reintroducing the dependence when formality is required (such as in the statements of theorems). In particular, we will not write probability and expectation operators ($\mathbb{P}$ and $\mathbb{E}$) with explicit dependence on n; this also applies to the functions $\rho(\cdot), \bar{\mu}_{k,\text{raw}}(\cdot), \bar{\mu}_{k,\text{cent}}(\cdot)$, and $\sigma_k(\cdot)$. The following assumption, constraint, and condition completes the description of our most general asymptotic setup.[10]

Assumption D.8 (General moment conditions). *Define*

$$L_{\text{fourth}}^{(n)} := \sup_{t \in [m(n)]_0 : \alpha_t^{(n)} > 0} \bar{\mu}_{4,\text{raw}}(\mathcal{R}_t^{(n)}),$$

$$L_{\text{invvar}}^{(n)} := \sup_{t \in [m(n)]_0 : \alpha_t^{(n)} > 0} \sigma^{-2}(\mathcal{R}_t^{(n)}).$$

Then $L_{\text{fourth}}^{(n)}$ and $L_{\text{invvar}}^{(n)}$ are bounded in probability.

Constraint D.9 (Tail error restriction). *Fix $\alpha_{\min} \in (0, 1)$. Algorithm 1 satisfies the following for all n almost surely:*

$$\text{If } \alpha_t^{(n)} > 0, \text{ then } \alpha_t^{(n)} \geq \alpha_{\min}.$$

Condition D.10 (Diverging sample size). *Define*

$$N_{\min}^{(n)} := \min\{n_t^{(n)} \text{ for } t = 0, ..., m(n) : \alpha_t^{(n)} > 0\}.$$

Then $N_{\min}^{(n)} \xrightarrow{p} \infty$.

Recall that Constraint D.9 implies that there exists a constant K_{reg} such that

$$|\{t = 0, ..., m(n) : \alpha_t^{(n)} > 0\}| \leq K_{\text{reg}} \text{ a.s. for all } n. \tag{21}$$

---

[10]A word on nomenclature: we have designated the first of these to be an assumption as what it imposes is essentially unverifiable. The second is a constraint since the analyst can satisfy it by choice. The last requires that our sample sizes are sufficiently large; as this is not exactly unverifiable, nor exactly satisfiable by choice, we call it a condition.

For instance, it suffices to let K_{reg} be such that $1 - (1 - \alpha_{\min})^{K_{\text{reg}}} \geq \alpha$. For convenience, we will simply treat the redundant Equation (21) as a part of Constraint D.9 throughout the appendix, and we will refer to it as the "bounded number of tested regions" constraint. In fact, Constraint D.9 is not exactly required to prove validity of the oracle test (Proposition D.11 below); rather, proofs for the oracle test only rely on Equation (21). However, Constraint D.9 is required for subsequent results concerning the empirical tests.

We are finally prepared to state the results we prove in this section.

Proposition D.11 (Error control for oracle test)**.** *When the sample size is n, let $\mathcal{R}_\tau^{(n)}$ be the region produced by Algorithm 1 using tests of the form $\phi_t^{(n)} := \mathbb{1}\{T_t^{(n)} > C_t^{(n)}\}$ as specified by Definition D.5. Then under Assumption D.8, Constraint D.9, and Condition D.10,*

$$\limsup_{n\to\infty} \mathbb{P}\big(\mu(\mathcal{R}_\tau^{(n)}) \leq 0\big) \leq \alpha.$$

A proof is given in Appendix D.4.4. Of course, the oracle test cannot be realized in practice since it depends on the unknown variances V_t. The next lemma connects the oracle test to the empirical test.

Lemma D.12 (Oracle/empirical algorithm convergence.)**.** *Let $\mathcal{R}_{\tau_{\text{oracle}}}^{(n)}$ be the region outputted by Algorithm 1 using the tests defined in Definition D.5. Let $\mathcal{R}_{\tau_{\text{empirical}}}^{(n)}$ be the region outputted by Algorithm 1 using the tests defined in Definition D.6 applied to the same dataset as the former. Note that in both instances we assume that we select the same $f(\cdot)$, c, and α_t in Algorithms 1 and 2. Then under Assumption D.8, Constraint D.9, and Condition D.10,*

$$\lim_{n\to\infty} \mathbb{P}\Big(\mathcal{R}_{\tau_{\text{oracle}}}^{(n)} = \mathcal{R}_{\tau_{\text{empirical}}}^{(n)}\Big) = 1.$$

A proof is given in Appendix D.4.5. Hence, the empirical test inherits error control from the oracle test.

Theorem D.13 (Error control for empirical test)**.** *When the sample size is n, let $\mathcal{R}_\tau^{(n)}$ be the region produced by Algorithm 1 using tests of the form $\hat{\phi}_t^{(n)} := \mathbb{1}\{T_t^{(n)} > \hat{C}_t^{(n)}\}$ as defined in Definition D.6. Then under Assumption D.8, Constraint D.9, and Condition D.10,*

$$\limsup_{n\to\infty} \mathbb{P}\big(\mu(\mathcal{R}_\tau^{(n)}) \leq 0\big) \leq \alpha.$$

A proof is given in Appendix D.4.6. Finally, we reiterate how this connects to the test and guarantee described in Section 3.6. As noted in Appendix D.4.1, the test described in Definition D.6 almost surely reports the same rejections as the test described in Definition 3.10 as long as Constraint 3.12 is satisfied. Moreover, we will show in Appendix D.4.7 that Assumption 3.11 and Constraint 3.12 imply Assumption D.8, Constraint 3.12, and Condition D.10. Hence, the guarantee of Theorem D.13 is more general than that of Theorem 3.13.

In the next section, we establish some intermediate properties that will be useful for proving the above results.

D.4.3 Implications of Assumption D.8, Constraint D.9, and Condition D.10

We state three properties that are implied by Assumption D.8, Constraint D.9, and Condition D.10, which will be helpful in deriving the asymptotic validity of the oracle and empirical tests for subgroup means. Note that Properties D.15–D.16 will not be used until Appendix E. In the definitions below, we take the convention that the supremum over an empty set is 0.

Property D.14 (Uniformly consistent approximations). *Define*

$$
L_{\text{approx}}^{(n)} := \sup_{t \in [m(n)]_0 : \alpha_t^{(n)} > 0} \sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n_t^{(n)}}}{\sqrt{V_t^{(n)}}} \left(\bar{\mu}_t^{(n)} - \mu(\mathcal{R}_t^{(n)}) \right) \le x \middle| \mathcal{F}_t^{(n)} \right) - \Phi(x) \right|.
$$

Then $L_{\text{approx}}^{(n)} \xrightarrow{p} 0$.

Property D.15 (Variance ratio bound). *Define*

$$
L_{\text{ratio}}^{(n)} := \sup_{s,t \in [m(n)]_0 : \alpha_s^{(n)}, \alpha_t^{(n)} > 0} \frac{n_s^{(n)} V_s^{(n)}}{n_t^{(n)} V_t^{(n)}} \cdot \mathbb{1}\{p_{\text{prop}} \le n_s/n_t \le p_{\text{prop}}^{-1}\}
$$

Then $L_{\text{ratio}}^{(n)}$ is bounded in probability.

Property D.16 (Uniformly consistent variance estimates). *Define*

$$
L_{\text{var}}^{(n)} := \sup_{t \in [m(n)]_0 : \alpha_t^{(n)} > 0} \left| \frac{\hat{V}_t^{(n)}}{V_t^{(n)}} - 1 \right|.
$$

Then $L_{\text{var}}^{(n)} \xrightarrow{p} 0$.

Now we state the lemma.

Lemma D.17 (Implied convergence properties). *Assumption D.8, Constraint D.9, and Condition D.10 imply Properties D.14, D.15, and D.16.*

A proof is given in Appendix E.1. With these, we may establish some results.

D.4.4 Proof of Proposition D.11

Proof. The strategy we take will be nearly identical to the proof of Theorem 3.9, except that we must keep track of an additional, vanishing error term. As a reminder, the conditions of the proposition imply, via Lemma D.17, that we have Property D.14. It suffices to show that that we satisfy the conditions of the asymptotic abstract error control result (Corollary D.4).

Let $(\mathcal{R}_t, \alpha_t, \mathcal{F}_t)_{t=0}^m$ be defined as in Algorithm 1 and note that we are suppressing dependence on n. Let $\mathcal{H}_t := \mathbb{1}\{\mu(\mathcal{R}_t) > 0\}$, and let ϕ_t be defined as in the proposition statement.

To apply Corollary D.4, we will let $\mathcal{G}_t := \mathcal{F}_t$. Define the following random quantities:

$$\tilde{M}_t := \min_{s<t}\left\{\frac{C_s - \Delta_{s,t}}{v_{s,t}}\right\}, \qquad\qquad\qquad\text{(true trunc. level)}$$

$$W_t := \frac{\sqrt{n_t}}{\sqrt{V_t}}\mu(\mathcal{R}_t), \qquad\qquad\qquad\qquad\text{(rescaled mean)}$$

$$Z_t := T_t - W_t = \frac{\sqrt{n_t}}{\sqrt{V_t}}(\bar{\mu}_t - \mu(\mathcal{R}_t)), \qquad\qquad\text{(Z-stat.)}$$

$$F_t(x;y) := \mathbb{P}(Z_t \le x | \mathcal{F}_t, Z_t \le y), \qquad\qquad\text{(trunc. dist. of Z-stat.)}$$

$$\delta_t := \left(\sup_{x\ge 0, y\in\mathbb{R}}|F_t(x;y) - \Phi(x;y)|\right)\cdot\mathbb{1}\{\alpha_t > 0\}, \quad\text{(normal approx. error)}$$

$$L_{\text{trunc}} := \sup_{t\in[m]_0:\alpha_t>0}\sup_{x\ge 0, y\in\mathbb{R}}|F_t(x;y) - \Phi(x;y)|. \qquad\text{(max approx. error)}$$

First, we show that

$$\mathbb{P}(T_t \le C_t \mid \sigma(\mathcal{F}_t, (\phi_s)_{s<t})) \ge 1 - \alpha_t - \delta_t \quad\text{conditional on}\quad \{\mathcal{H}_t = 0 \text{ and } \max_{s<t}\phi_s = 0\}.$$

We rewrite the left-hand side. Implicitly condition on $\{\mathcal{H}_t = 0 \text{ and } \max_{s<t}\phi_s = 0\}$ throughout. Making use of the equivalence described in Equation (20), the first few steps of the expansion are identical to the analogous expansion in the proof of Theorem 3.9, so we can skip ahead to deduce the first equality below, and then continue expanding:

$$\begin{aligned}
\mathbb{P}(T_t \le C_t \mid \sigma(\mathcal{F}_t, (\phi_s)_{s<t})) &= \mathbb{P}(T_t \le C_t \mid \mathcal{F}_t, T_t \le \tilde{M}_t)\\
&= \mathbb{P}(T_t - W_t \le C_t - W_t \mid \mathcal{F}_t, T_t - W_t \le \tilde{M}_t - W_t)\\
&= \mathbb{P}(Z_t \le C_t - W_t \mid \mathcal{F}_t, Z_t \le \tilde{M}_t - W_t).
\end{aligned}$$

Next, note that since C_t, $\tilde{M}_t$, and W_t are $\mathcal{F}_t$-measurable, we may make the formal substitution

$$F_t(C_t - W_t; \tilde{M}_t - W_t) = \mathbb{P}(Z_t \le C_t - W_t \mid \mathcal{F}_t, Z_t \le \tilde{M}_t - W_t),$$

and hence

$$\begin{aligned}
&|\mathbb{P}(Z_t \le C_t - W_t \mid \mathcal{F}_t, Z_t \le \tilde{M}_t - W_t) - \Phi(C_t - W_t; \tilde{M}_t - W_t)|\\
=&|F_t(C_t - W_t; \tilde{M}_t - W_t) - \Phi(C_t - W_t; \tilde{M}_t - W_t)|\\
\le& \sup_{x\ge 0, y\in\mathbb{R}}|F_t(x;y) - \Phi(x;y)| \cdot \mathbb{1}\{\alpha_t > 0\}\\
=&\delta_t.
\end{aligned}$$

To reach the third line of the above, we use the fact that $C_t \ge 0$ almost surely and that $W_t \le 0$, so that $C_t - W_t \ge 0$ almost surely (conditional on $\mathcal{H}_t = 0$). We also use the fact that when $\alpha_t = 0$ then $C_t \ge M_t \ge \tilde{M}_t$ almost surely and hence the difference in the second

line is 0. Putting the above together,

$$
\begin{aligned}
&\mathbb{P}(T_t \leq C_t \mid \sigma(\mathcal{F}_t, (\phi_s)_{s<t}))\\
&=\Phi(C_t - W_t; \tilde{M}_t - W_t)\\
&\quad + \mathbb{P}(Z_t \leq C_t - W_t \mid \mathcal{F}_t, Z_t \leq \tilde{M}_t - W_t) - \Phi(C_t - W_t; \tilde{M}_t - W_t)\\
&\geq \Phi(C_t - W_t; \tilde{M}_t - W_t)\\
&\quad - |\mathbb{P}(Z_t \leq C_t - W_t \mid \mathcal{F}_t, Z_t \leq \tilde{M}_t - W_t) - \Phi(C_t - W_t; \tilde{M}_t - W_t)|\\
&\geq \Phi(C_t - W_t; \tilde{M}_t - W_t) - \delta_t\\
&\geq \Phi(C_t; \tilde{M}_t) - \delta_t\\
&\geq \Phi(C_t; M_t) - \delta_t\\
&\geq 1 - \alpha_t - \delta_t
\end{aligned}
$$

where to reach the third to last line we recall that in this case $W_t \leq 0$ almost surely, so we may apply a stochastic dominance lemma for truncated normals (Lemma F.8). The second to last line follows from the fact that $\Phi(x; y)$ is decreasing in its second argument and $\tilde{M}_t \leq M_t$. The last line follows since $\Phi(C_t; M_t) \geq 1 - \alpha_t$ by construction.

To apply Corollary D.4, it remains to be shown that the sum of the errors vanishes in L^1. Note that $|\delta_t| \leq L_{\text{trunc}}$ almost surely for all t, and so

$$
\begin{aligned}
\sum_{t=0}^{m} |\delta_t| &= \sum_{t=0}^{m} |\delta_t| \cdot \mathbb{1}\{\alpha_t > 0\}\\
&\leq \sum_{t=0}^{m} L_{\text{trunc}} \cdot \mathbb{1}\{\alpha_t > 0\}\\
&= L_{\text{trunc}} \cdot \sum_{t=0}^{m} \mathbb{1}\{\alpha_t > 0\}\\
&\leq L_{\text{trunc}} \cdot K_{\text{reg}}\\
&\xrightarrow{p} 0.
\end{aligned}
$$

In the above, we applied Constraint 3.12 to reach the second to last line. Furthermore, Property D.14 implies that L_{trunc} converges to 0 in probability (Lemma F.9), which allows us to conclude the last line. Since $L_{\text{trunc}} \leq 1$ almost surely, by dominated convergence we can conclude that $\sum_{t=0}^{m} |\delta_t| \xrightarrow{L^1} 0$. Thus, we may apply Corollary D.4, and the proposition follows. $\qquad\square$

D.4.5 Proof of Lemma D.12

Before we proceed to the proof of Lemma D.12, we need to show that certain quantities converge in probability. Define the following quantities that we will use throughout:

$$d_t := C_t - \sqrt{\frac{n_t}{V_t}} \cdot \mu(\mathcal{R}_t), \qquad \text{(critical value minus normalized mean)}$$

$$\hat{d}_t := \sqrt{\frac{\hat{V}_t}{V_t}} \cdot \hat{C}_t - \sqrt{\frac{n_t}{V_t}} \cdot \mu(\mathcal{R}_t), \qquad \text{(empirical variant of } d_t) \qquad (22)$$

$$Z_t := T_t - \sqrt{\frac{n_t}{V_t}} \cdot \mu(\mathcal{R}_t) \qquad \qquad (Z\text{-stat.})$$

and note that equivalently,

$$Z_t = \frac{1}{\sqrt{V_t} \cdot \sqrt{n_t}} \sum_{i=1}^{n} (Y_i - \mu(\mathcal{R}_t)) \cdot \mathbb{1}\{X_i \in \mathcal{R}_t\}.$$

Lemma D.18 (Uniform convergence of shifted critical values)**.** *Under Assumption D.8, Constraint D.9, and Condition D.10, we have*

$$\sup_{t \in [m(n)]_0 : \alpha_t^{(n)} > 0} |d_t^{(n)} - \hat{d}_t^{(n)}| \xrightarrow{p} 0$$

where here we take the convention that the supremum over an empty set is 0.

A proof is given in Appendix E.2. We now prove the lemma.

Proof of Lemma D.12. Since the sequence of regions produced by both the oracle and empirical algorithm are assumed to be the same (as they make the same decisions regarding $f(\cdot)$ and c in Algorithm 2), it suffices to show that their first rejections coincide with probability going to 1. In particular, it suffices to show that the probability that $\phi_t \neq \hat{\phi}_t$ for some $t \in [m]_0$ vanishes.

Note that

$$\phi_t \neq \hat{\phi}_t \text{ for some } t \in [m]_0$$
$$\implies \mathbb{1}\{T_t \leq C_t\} \neq \mathbb{1}\{\hat{T}_t \leq \hat{C}_t\} \text{ and } \alpha_t > 0 \text{ for some } t \in [m]_0.$$

Next, note that since $\sqrt{\frac{\hat{V}_t}{V_t}} \cdot \hat{T}_t = T_t$ for all t, we have

$$\mathbb{1}\{T_t \leq C_t\} \neq \mathbb{1}\{\hat{T}_t \leq \hat{C}_t\} \iff \min\left\{C_t, \sqrt{\frac{\hat{V}_t}{V_t}} \cdot \hat{C}_t\right\} < T_t \leq \max\left\{C_t, \sqrt{\frac{\hat{V}_t}{V_t}} \cdot \hat{C}_t\right\}$$

$$\iff \min\{d_t, \hat{d}_t\} < Z_t \leq \max\{d_t, \hat{d}_t\}$$

where we recall the definition of d_t, $\hat{d}_t$, and Z_t from Equation (22). For a fixed $\delta > 0$ that we will specify later, let $E := \{\sup_{t \in [m]_0 : \alpha_t > 0} |d_t - \hat{d}_t| < \delta\}$. Then for each $t \in [m]_0$,

$$\mathbb{P}(\min\{d_t, \hat{d}_t\} < Z_t \leq \max\{d_t, \hat{d}_t\}, \alpha_t > 0, E) \leq \mathbb{P}(d_t - \delta < Z_t \leq d_t + \delta, \alpha_t > 0).$$

Now, conditional on $\{\alpha_t > 0\}$ and recalling that α_t is $\mathcal{F}_t$-measurable,

$$\mathbb{P}(d_t - \delta < Z_t \le d_t + \delta \mid \mathcal{F}_t)$$
$$=\mathbb{P}(Z_t \le d_t + \delta \mid \mathcal{F}_t) - \mathbb{P}(Z_t \le d_t - \delta \mid \mathcal{F}_t)$$
$$=\mathbb{P}(Z_t \le d_t + \delta \mid \mathcal{F}_t) - \Phi(d_t + \delta) - \mathbb{P}(Z_t \le d_t - \delta \mid \mathcal{F}_t) + \Phi(d_t - \delta)$$
$$\quad + \Phi(d_t + \delta) - \Phi(d_t - \delta)$$
$$\le |\mathbb{P}(Z_t \le d_t + \delta \mid \mathcal{F}_t) - \Phi(d_t + \delta)| + |\mathbb{P}(Z_t \le d_t - \delta \mid \mathcal{F}_t) - \Phi(d_t - \delta)|$$
$$\quad + |\Phi(d_t + \delta) - \Phi(d_t - \delta)|$$
$$\le \sup_{x \in \mathbb{R}} |\mathbb{P}(Z_t \le x \mid \mathcal{F}_t) - \Phi(x)| + \sup_{x \in \mathbb{R}} |\mathbb{P}(Z_t \le x \mid \mathcal{F}_t) - \Phi(x)|$$
$$\quad + |\Phi(d_t + \delta) - \Phi(d_t - \delta)|$$
$$\le 2L_{\text{approx}} + 2\delta.$$

To reach the fifth line of the above, we use the fact that d_t is $\mathcal{F}_t$-measurable and is hence a constant conditional on $\mathcal{F}_t$. To reach the last line, we recall the definition on L_{approx} from Property D.14, recalling that it is implied by Lemma D.17. We also use the fact that Φ is 1-Lipschitz to bound $|\Phi(d_t + \delta) - \Phi(d_t - \delta)| \le 2\delta$. Thus for all $t \in [m]_0$,

$$\mathbb{P}(\min\{d_t, \hat{d}_t\} < Z_t \le \max\{d_t, \hat{d}_t\}, \alpha_t > 0, E)$$
$$\le \mathbb{E}[\mathbb{P}(d_t - \delta < Z_t \le d_t + \delta, \alpha_t > 0 \mid \mathcal{F}_t)]$$
$$=\mathbb{E}[\mathbb{P}(d_t - \delta < Z_t \le d_t + \delta \mid \mathcal{F}_t)\mathbb{1}\{\alpha_t > 0\}] \tag{23}$$
$$\le \mathbb{E}[(2L_{\text{approx}} + 2\delta) \cdot \mathbb{1}\{\alpha_t > 0\}].$$

Now let $\epsilon > 0$. Choose $\delta = \epsilon/(8K_{\text{reg}})$. Since $L_{\text{approx}} \xrightarrow{p} 0$ by Property D.14 and $L_{\text{approx}} \le 1$ almost surely, then dominated convergence implies that $\lim_{n \to \infty} \mathbb{E}[L_{\text{approx}}] = 0$. In particular, we may choose N_1 such that $\mathbb{E}[L_{\text{approx}}] < \epsilon/(8K_{\text{reg}})$ for all $n > N_1$. By Lemma D.18, we can choose N_2 such that $\mathbb{P}(E) > 1 - \epsilon/2$ for all $n > N_2$. Then for all $n > \max\{N_1, N_2\}$,

$$\mathbb{P}(\phi_t \ne \hat{\phi}_t \text{ for some } t \in [m]_0)$$
$$\le \mathbb{P}(\min\{d_t, \hat{d}_t\} < Z_t \le \max\{d_t, \hat{d}_t\} \text{ and } \alpha_t > 0 \text{ for some } t \in [m]_0)$$
$$\le \mathbb{P}(E^{\mathsf{c}}) + \mathbb{P}(\min\{d_t, \hat{d}_t\} < Z_t \le \max\{d_t, \hat{d}_t\} \text{ and } \alpha_t > 0 \text{ for some } t \in [m]_0, E)$$
$$\le \epsilon/2 + \sum_{t=0}^{m} \mathbb{P}(\min\{d_t, \hat{d}_t\} < Z_t \le \max\{d_t, \hat{d}_t\} \text{ and } \alpha_t > 0, E)$$
$$\le \epsilon/2 + \sum_{t=0}^{m} \mathbb{E}[(2L_{\text{approx}} + 2\delta) \cdot \mathbb{1}\{\alpha_t > 0\}]$$
$$=\epsilon/2 + \mathbb{E}\left[(2L_{\text{approx}} + 2\delta) \cdot \sum_{t=0}^{m} \mathbb{1}\{\alpha_t > 0\}\right]$$
$$\le \epsilon/2 + 2K_{\text{reg}}\mathbb{E}[L_{\text{approx}}] + 2K_{\text{reg}}\delta$$
$$\le \epsilon/2 + \epsilon/4 + \epsilon/4$$
$$=\epsilon.$$

Note that we have applied Constraint D.9 to conclude that $\sum_{t=0}^{m} \mathbb{1}\{\alpha_t > 0\} \le K_{\text{reg}}$ almost surely. As discussed before, this establishes the result. $\qquad \square$

D.4.6 Proof of Theorem D.13

Proof. Let τ_{oracle} and $\tau_{\mathrm{empirical}}$ be defined as in Lemma D.12. Then

$$
\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}(\mu(\mathcal{R}_{\tau_{\mathrm{empirical}}}) \leq 0) \\
&\leq \limsup_{n\to\infty} \mathbb{P}(\mu(\mathcal{R}_{\tau_{\mathrm{empirical}}}) \leq 0, \mathcal{R}_{\tau_{\mathrm{oracle}}} = \mathcal{R}_{\tau_{\mathrm{empirical}}}) + \mathbb{P}(\mathcal{R}_{\tau_{\mathrm{oracle}}} \neq \mathcal{R}_{\tau_{\mathrm{empirical}}}) \\
&= \limsup_{n\to\infty} \mathbb{P}(\mu(\mathcal{R}_{\tau_{\mathrm{empirical}}}) \leq 0, \mathcal{R}_{\tau_{\mathrm{oracle}}} = \mathcal{R}_{\tau_{\mathrm{empirical}}}) \\
&= \limsup_{n\to\infty} \mathbb{P}(\mu(\mathcal{R}_{\tau_{\mathrm{oracle}}}) \leq 0, \mathcal{R}_{\tau_{\mathrm{oracle}}} = \mathcal{R}_{\tau_{\mathrm{empirical}}}) \\
&\leq \limsup_{n\to\infty} \mathbb{P}(\mu(\mathcal{R}_{\tau_{\mathrm{oracle}}}) \leq 0) \\
&\leq \alpha
\end{aligned}
$$

where the third line follows from Lemma D.12 and the last line follows from Proposition D.11.

$\square$

D.4.7 Sufficient primitive conditions

We describe primitive sufficient conditions where our theory applies.

Pointwise validity. First, we consider the setting described in Section 3.6. The pointwise asymptotics described in that section can be embedded into our triangular array setting by letting each row be drawn i.i.d. from the same distribution. Also, Constraint 3.12 directly implies Constraint D.9 and Condition D.10. It remains to show that Assumption 3.11 and Constraint 3.12 imply Assumption D.8. For the first part,

$$
\sup_{t:\alpha_t>0} \bar{\mu}_{4,\mathrm{raw}}(\mathcal{R}_t) \leq \sup_{t:\alpha_t>0} \frac{\mathbb{E}[|Y|^4]}{\rho(\mathcal{R}_t)} \leq \mathbb{E}[|Y|^4] \cdot \sup_{t:n_t/n\geq p_{\mathrm{prop}}} \rho(\mathcal{R}_t)^{-1}.
$$

The expectation of the supremum on the right-hand side is bounded by a universal constant that only depends on p_{prop} (Corollary E.12). Therefore, introducing dependence on n, the right-hand side is bounded in probability by Markov's inequality. As for the inverse variance part of Assumption D.8, this follows directly from Lemma E.13.

Uniform validity. Suppose that we still enforce Constraint 3.12 but now we allow the distributions to vary with n in the triangular array. If we suppose that there exists constants $L, B > 0$ such that

$$
\mathbb{E}[|Y^{(n)}|^4] \leq L \quad \text{and} \quad \mathrm{Var}(Y^{(n)} \mid X^{(n)}) \geq B \text{ a.s.}
$$

for all n, then the fourth moment part of Assumption D.8 follows from an identical calculation as in the pointwise case, while the inverse variance part follows immediately from the fact that all region inverse variances are bounded almost surely.

Relaxing p_{prop} constraint. Lastly, suppose that we enforce Constraint D.9 and Condition D.10 directly. This only requires that the minimum sample size diverges to ∞ at any

rate. Then a variety of weak distributional assumptions suffice to imply Assumption D.8. For instance, simply requiring that for some $M, B > 0$,

$$|Y^{(n)}| \leq M \quad \text{and} \quad \text{Var}(Y^{(n)} \mid X^{(n)}) \geq B \text{ a.s.}$$

for all n suffices. A weaker and more sophisticated sufficient condition is that for some $M, L, B > 0$,

$$|\mathbb{E}[Y^{(n)} \mid X^{(n)}]| \leq M,$$
$$\mathbb{E}[(Y^{(n)} - \mathbb{E}[Y^{(n)} \mid X^{(n)}])^4 \mid X^{(n)}] \leq L,$$
$$\text{Var}(Y^{(n)} \mid X^{(n)}) \geq B$$

almost surely for all n. Note that we have relaxed p_{prop} in the sense of getting rid of Constraint 3.12, but p_{prop} still appears in the definitions of the tests, e.g. Definition D.6.

D.4.8 Discussion of distal truncation information

Our generalized test ignores distal truncation information (distal in the sense that once the sample size of an earlier test becomes incomparable to the current sample size, that truncation information is ignored). This was designed in order to overcome some technical difficulties in our proofs; however, we expect this choice to have a limited effect in practice. To get a heuristic sense for the features driving the asymptotic behavior, consider the following calculations. Let (X, Y) be a fixed distribution and let $\mathcal{R}^{(1)}, \mathcal{R}^{(2)}, \ldots \subseteq \mathcal{X}$ be a fixed sequence of regions. Let $\rho_n := \mathbb{P}(X \in \mathcal{R}^{(n)})$. We suppose that $\rho_n \to 0$ while $n\rho_n \to \infty$. Define $k_n := \sum_{i=1}^n \mathbb{1}\{X_i \in \mathcal{R}^{(n)}\}$ and note that $k_n \overset{p}{\to} \infty$ since

$$\mathbb{E}[k_n] = n\rho_n, \quad \text{Var}(k_n) = n\rho_n(1 - \rho_n) \leq n$$

so that for all $\delta > 0$ Chebyshev's inequality yields

$$\mathbb{P}(k_n \leq n\rho_n - \delta\sqrt{n}) \leq \frac{1}{\delta^2} \quad \text{for all } n.$$

Then we simply note that $n\rho_n - \delta\sqrt{n} \to \infty$ for any $\delta > 0$. Also, k_n/n concentrates around ρ_n, and since the latter is going to 0 we have $k_n/n \overset{p}{\to} 0$.

Now define $\theta_0 := \mu(\mathcal{X})$, $\sigma_0^2 := \text{Var}(Y)$, $\theta_n := \mu(\mathcal{R}^{(n)})$, and $\sigma_n^2 := \text{Var}(Y \mid X \in \mathcal{R}^{(n)})$. Then define

$$Z_n := \frac{1}{\sigma_0\sqrt{n}} \sum_{i=1}^n (Y_i - \theta_0) \quad \text{and} \quad W_n := \frac{1}{\sigma_n\sqrt{k_n}} \sum_{i:X_i \in \mathcal{R}^{(n)}} (Y_i - \theta_n).$$

For simplicity suppose that Y is bounded and $\text{Var}(Y \mid X) \geq b$ a.s. for some $b > 0$; then both of the above converge to standard normal distributions. More crucially, we can show that Z_n and W_n are asymptotically independent, which we do at the end of this section. This result suggests why we do not expect to lose much from ignoring very distal truncation information in Definitions D.5 and D.6. Tests that utilize comparable sample sizes pass along their truncation information, but for tests with incomparable sample sizes the truncation information is asymptotically irrelevant. Intuitively, this is because W_n contributes a

vanishingly small fluctuation to Z_n. To say something slightly more concrete: Suppose that $\theta_0 = 0$ and $\mu_{\mathrm{cut}} = 0$, and that we run the oracle version of Algorithm 1 to test $H_0 : \theta_0 \leq 0$, then immediately chisel to $\mathcal{R}^{(n)}$ and test θ_n. If we do not reject at the first stage, then we pass along the information that $Z_n \leq C_0$ to the test of $H_0 : \theta_n \leq 0$, which is based on the distribution of W_n. But conditioning on $Z_n \leq C_0$ asymptotically does not affect the distribution of W_n, and thus in principle we do not sacrifice much by ignoring the truncation information, as we do in Definitions D.5 and D.6 once k_n becomes sufficiently small relative to n.

Proof that Z_n and W_n are asymptotically independent. Without loss of generality we may assume $\theta_0 = 0$ by subtracting off θ_0 from Y. Then by straightforward algebra we may write

$$
Z_n = \underbrace{\frac{\sigma_n \sqrt{k_n}}{\sigma_0 \sqrt{n}}\left(W_n + \frac{\sqrt{k_n}}{\sigma_n}\theta_n \right)}_{\delta_n} + \frac{\bar{\sigma}_n \sqrt{n - k_n}}{\sigma_0 \sqrt{n}} \times \underbrace{\frac{1}{\bar{\sigma}_n \sqrt{n - k_n}} \sum_{i : X_i \notin \mathcal{R}^{(n)}} (Y_i - \bar{\theta}_n)}_{Z_n^*} + \underbrace{\frac{n - k_n}{\sigma_0 \sqrt{n}}\bar{\theta}_n}_{\epsilon_n}
$$

where $\bar{\theta}_n := \mathbb{E}[Y \mid X \notin \mathcal{R}^{(n)}]$ and $\bar{\sigma}_n^2 := \mathrm{Var}(Y \mid X \notin \mathcal{R}^{(n)})$. Conditionally on W_n and k_n, we have that δ_n and ϵ_n are constants while Z_n^* is approximately standard normal. In particular, the Berry-Esseen theorem gives

$$
\mathbb{P}(Z_n \leq z \mid W_n, k_n) = \Phi\left((z - \delta_n - \epsilon_n) \cdot \frac{\sigma_0 \sqrt{n}}{\bar{\sigma}_n \sqrt{n - k_n}} \right) + \beta_n
$$

where β_n is the approximation error, a random variable such that $|\beta_n| \leq K/\sqrt{n - k_n}$ almost surely for some universal constant K (since Y is bounded with lower bounded conditional variance). Then $\beta_n \to 0$ in probability. Furthermore, $\bar{\sigma}_n \to \sigma_0$ and $\sqrt{n/(n - k_n)} \xrightarrow{p} 1$. It remains to show that $\delta_n + \epsilon_n \xrightarrow{p} 0$; once we've shown this, it follows by Slutsky's theorem and the continuous mapping theorem that

$$
\mathbb{P}(Z_n \leq z \mid W_n, k_n) \xrightarrow{p} \Phi(z)
$$

and hence Z_n is asymptotically independent of W_n and k_n.

To see that $\delta_n + \epsilon_n \xrightarrow{p} 0$, we expand

$$
\begin{aligned}
\delta_n + \epsilon_n &= \frac{\sigma_n \sqrt{k_n}}{\sigma_0 \sqrt{n}} W_n + \frac{k_n}{\sigma_0 \sqrt{n}}\theta_n + \frac{n - k_n}{\sigma_0 \sqrt{n}}\bar{\theta}_n \\
&= \frac{\sigma_n \sqrt{k_n}}{\sigma_0 \sqrt{n}} W_n + \frac{1}{\sigma_0} \cdot \sqrt{n} \cdot \frac{k_n}{n}\theta_n + \frac{1}{\sigma_0} \cdot \sqrt{n} \cdot \frac{n - k_n}{n}\bar{\theta}_n.
\end{aligned}
$$

The first term is easy to deal with since W_n is of constant order, σ_n/σ_0 is bounded, and $\sqrt{k_n/n} \xrightarrow{p} 0$. To deal with the last two terms, we first note that

$$
\begin{aligned}
\rho_n \theta_n + (1 - \rho_n)\bar{\theta}_n &= \mathbb{P}(X \in \mathcal{R}^{(n)})\mathbb{E}[Y \mid X \in \mathcal{R}^{(n)}] + \mathbb{P}(X \notin \mathcal{R}^{(n)})\mathbb{E}[Y \mid X \notin \mathcal{R}^{(n)}] \\
&= \mathbb{E}[Y] \\
&= 0.
\end{aligned}
$$

Then

$$\frac{1}{\sigma_0} \cdot \sqrt{n} \cdot \frac{k_n}{n} \theta_n + \frac{1}{\sigma_0} \cdot \sqrt{n} \cdot \frac{n - k_n}{n} \bar{\theta}_n = \frac{\theta_n}{\sigma_0} \cdot \underbrace{\sqrt{n}\left(\frac{k_n}{n} - \rho_n\right)}_{A_1} + \frac{\bar{\theta}_n}{\sigma_0} \cdot \underbrace{\sqrt{n}\left(\frac{n - k_n}{n} - (1 - \rho_n)\right)}_{A_2}.$$

Now θ_n/σ_0 and $\bar{\theta}_n/\sigma_0$ are bounded, so it suffices to show that A_1 and A_2 converge to 0 in probability. Note that $A_1 = -A_2$, that $\mathbb{E}[A_1] = 0$, and that

$$\mathrm{Var}(A_1) = n \cdot \frac{\rho_n(1 - \rho_n)}{n} = \rho_n(1 - \rho_n) \to 0.$$

Thus, A_1 concentrates around its mean and hence both A_1 and A_2 converge to 0 in probability. Combining this with the above yields $\delta_n + \epsilon_n \xrightarrow{P} 0$ as desired. $\qquad\square$

D.5 Validity results for AIPW

The IPW estimator is less efficient than the AIPW estimator, which reduces the variance of the ATE estimate by eliminating some nuisance variation. Moreover, in an observational study, it is often necessary to estimate the propensity scores in order to conduct inference on the ATE. Here, we describe how one can make use of the augmented IPW (AIPW) estimator in our framework. We describe the general construction and rigorously prove validity for the setting where the propensities are known (i.e. AIPW estimation in RCTs). We anticipate that the proof for the general case when propensities are also estimated from the data is similar, though we leave it for future work.

D.5.1 Setup

We use slightly different notation from the main text. Let $(Y_i'(1), Y_i'(0))$ denote potential outcomes for unit i and suppose we observe $(X_i, W_i, Y_i')_{i=1}^n$ where $Y_i' := W_i Y_i'(1) + (1 - W_i)Y_i'(0)$ is the observed outcome, W_i is a binary exposure indicator, and X_i is a vector of pre-exposure covariates belonging to $\mathcal{X}$. If $g_1(\cdot)$ and $g_0(\cdot)$ are any functions and $e(x) := \mathbb{P}(W = 1 \mid X = x)$, then define the transformed outcomes

$$Y_{i,1} := g_1(X_i) + \frac{W_i(Y_i' - g_1(X_i))}{e(X_i)},$$
$$Y_{i,0} := g_0(X_i) - \frac{(1 - W_i)(Y_i' - g_0(X_i))}{1 - e(X_i)},$$

and $Y_i := Y_{i,1} - Y_{i,0}$. Then

$$\mu(\mathcal{R}) := \mathbb{E}[Y'(1) - Y'(0) \mid X \in \mathcal{R}] = \mathbb{E}[Y \mid X \in \mathcal{R}]$$

as long as the latter expectation exists. Under additional standard regularity/strong overlap assumptions, the sample mean of the transformed outcomes obeys a central limit theorem. The variance of $\tilde{Y}_i$ is minimized when $g_1(\cdot)$ and $g_0(\cdot)$ are equal to $\mathbb{E}[Y'(1) \mid X = x]$ and $\mathbb{E}[Y'(0) \mid X = x]$ respectively.

When the propensity $e(\cdot)$ is known and the functions $g_1(\cdot)$ and $g_0(\cdot)$ are fixed *a priori*, then we may simply run chiseling on the dataset $(\tilde{X}_i, Y_i)_{i=1}^n$ where $\tilde{X}_i := (X_i, W_i, Y_i')$, making sure to only chisel within the X-dimension so that regions are subsets of $\mathcal{X}$. We include W_i and Y_i' as part of $\tilde{X}_i$ so that the analyst at least has access to the original pre-transformed data. The theory we have established up to now for general outcomes fully applies to this procedure as long as the transformed outcome obeys the asymptotic conditions outlined in Section 3.6 or the more general ones described in Appendix D.4.

We move beyond the scope of the theory established up to now when $g_1(\cdot)$ and $g_0(\cdot)$, and possibly $e(\cdot)$, are estimated from the data. We propose, as is now common, to estimate them via cross-fitting. Let there be $K_{\mathrm{cv}} \geq 2$ folds. Let $\mathcal{I}_1, ..., \mathcal{I}_{K_{\mathrm{cv}}}$ be random, approximately even partitions of $[n]$. For $w = 0, 1$ and $j = 1, ..., K_{\mathrm{cv}}$, let $\hat{g}_w^{(j)}(\cdot)$ be an estimate of $\mathbb{E}[Y(w) \mid X = x]$ and $\hat{e}^{(j)}(\cdot)$ be an estimate of $\mathbb{P}(W = 1 \mid X = x)$ using the data in $(X_i, W_i, Y_i')_{i \notin \mathcal{I}_j}$. Define

$$\hat{Y}_{i,1} := \sum_{j=1}^{K_{\mathrm{cv}}} \left(\hat{g}_1^{(j)}(X_i) + \frac{W_i(Y_i' - \hat{g}_1^{(j)}(X_i))}{\hat{e}^{(j)}(X_i)} \right) \mathbb{1}\{i \in \mathcal{I}_j\},$$

$$\hat{Y}_{i,0} := \sum_{j=1}^{K_{\mathrm{cv}}} \left(\hat{g}_0^{(j)}(X_i) + \frac{(1 - W_i)(Y_i' - \hat{g}_0^{(j)}(X_i))}{1 - \hat{e}^{(j)}(X_i)} \right) \mathbb{1}\{i \in \mathcal{I}_j\},$$

and $\hat{Y}_i := \hat{Y}_{i,1} - \hat{Y}_{i,0}$. The *AIPW variant* of our procedure runs Algorithm 1 on $(X_i, W_i, Y_i')_{i=1}^n$ (only shrinking in the X dimension) while using $(\hat{Y}_i)_{i=1}^n$ to calculate the tests. In particular, no subset of $(\hat{Y}_i)_{i=1}^n$ is ever revealed to the analyst.

We will establish theory for the case when the propensity $e(\cdot)$ is known and used to construct $\hat{Y}_i$, i.e. $\hat{e}^{(j)}(\cdot) = e(\cdot)$ for $j = 1, ..., K_{\mathrm{cv}}$. This is the case in randomized experiments, where the role of the AIPW adjustment is to increase efficiency by eliminating nuisance variation via $\hat{g}_w^{(j)}(\cdot)$. In this case, we only need the following assumption on the learned nuisance functions. We introduce indexing by n to accomodate the triangular array asymptotic setup we described in Appendix D.4.2, but will subsequently drop it for clarity.

Assumption D.19 (Stable nuisance estimation). *There exist functions $g_{1,n}(\cdot)$ and $g_{0,n}(\cdot)$ such that*

$$L_{\mathrm{stable}}^{(n)} := \sup_{x \in \mathrm{supp}(X^{(n)})} |\hat{g}_{w,n}^{(j)}(x) - g_{w,n}(x)| \xrightarrow{p} 0$$

for all $w = 0, 1$ and $j = 1, ..., K_{\mathrm{cv}}$, where $\mathrm{supp}(X^{(n)})$ denotes the support of the random variable $X^{(n)}$.

In particular, Assumption D.19 does not require $g_w(\cdot)$ to be the actual CATE function $\mathbb{E}[Y(w) \mid X = x]$, though the closer it is the greater the efficiency gain. It only requires that the nuisance estimators stabilize in the sense of being coupled to some deterministic sequence, and is thus extremely mild and satisfied for reasonable low-dimensional models and well-regularized estimation procedures.

When the propensities $e(\cdot)$ must also be estimated via cross-fitting, we will at minimum need to additionally impose the usual rate conditions from semiparametric efficiency theory on the nuisance estimators. One version of these is to require that $\hat{e}^{(j)}(\cdot)$ is sup norm

consistent for $e(x)$ and that

$$\mathbb{E}\big[(\hat{g}_w^{(j)}(X) - g_w(X))^2\big]\mathbb{E}\big[(\hat{e}^{(j)}(X) - e(X))^2\big] = o\left(\frac{1}{n}\right) \quad \text{for } w \in \{0,1\}, j \in \{1,...,K_{\text{cv}}\}.$$

See Chapter 3.1 on "Double machine learning" in Wager (2024a) for a discussion of these rate conditions and exemplar calculations utilizing them. For our purposes, some additional mild regularity conditions may be required. We leave an investigation to future work.

Remark D.20 (Translation invariance). *In practice, we recommend at least letting $\hat{g}_1^{(j)}(\cdot)$ and $\hat{g}_0^{(j)}(\cdot)$ be intercept models for each treatment group calibrated via cross-fitting, since the IPW estimator has the unappealing property of lacking translation invariance, while the AIPW with intercept model is asymptotically translation invariant. Of course, greater efficiency gains are possible by leveraging simple, low-dimensional models incorporating explanatory variables.*

D.5.2   Defining the AIPW-based tests

Let $(Y_i)_{i=1}^n$ and $(\hat{Y}_i)_{i=1}^n$ be defined as in Appendix D.5.1. Throughout, we will suppose that chiseling only shrinks in the X dimension; that is, all regions can be thought of subsets of $\mathcal{X}$. Throughout Appendix D.5 only, for $s \leq t$ we define

$$
\begin{aligned}
\hat{\mu}_t &:= \frac{1}{n_t} \sum_{i:X_i \in \mathcal{R}_t} \hat{Y}_i, \\
\hat{V}_t &:= \frac{1}{n_t} \sum_{i:X_i \in \mathcal{R}_t} (\hat{Y}_i - \hat{\mu}_t)^2 \\
\hat{\Delta}_{s,t} &:= \frac{1}{\sqrt{\hat{V}_s} \cdot \sqrt{n_s}} \sum_{i:X_i \in \mathcal{R}_s \setminus \mathcal{R}_t} \hat{Y}_i \\
\hat{v}_{s,t} &:= \sqrt{\frac{\hat{V}_t}{\hat{V}_s}}\sqrt{\frac{n_t}{n_s}}.
\end{aligned}
\tag{24}
$$

Note that compared to the definitions in Appendix D.4.1, $\hat{V}_t$, $\hat{\Delta}_{s,t}$, and $\hat{v}_{s,t}$ not only use the empirical variances but also $\hat{Y}_i$ in place of Y_i. The following formalizes the AIPW test.

Definition D.21. *(AIPW test statistics and critical values) Define $\hat{T}_t := \sqrt{n_t} \cdot \hat{V}_t^{-1/2}\hat{\mu}_t$. Also, recursively define*

$$\hat{M}_t := \min_{s<t:n_t/n_s \geq p_{\text{prop}}} \left\{\frac{\hat{C}_s - \hat{\Delta}_{s,t}}{\hat{v}_{s,t}}\right\} \quad and \quad \hat{C}_t := \max\{0, \Phi^{-1}(1 - \alpha_t; \hat{M}_t)\}$$

where the minimum is ∞ if the set is empty (i.e. when $t = 0$). The empirical tests are defined as $\hat{\phi}_t := \mathbb{1}\{\hat{T}_t > \hat{C}_t\}$.

D.5.3 Guarantees and proof outline

Recall the region moment functions from Equation (18), which we will also use here. We have aligned our notation so that Equation (18) should be imported exactly as is: for instance, $\bar{\mu}_{k,\mathrm{raw}}(\mathcal{R}) = \mathbb{E}[|Y|^k \mid X \in \mathcal{R}] \neq \mathbb{E}[|Y'|^k \mid X \in \mathcal{R}]$, which is an important distinction. The former refers to the moments of the oracle transformed outcome (defined in Appendix D.5.1), while the latter refers to moments of the raw outcome that the analyst sees. We can establish the following.

Theorem D.22 (Validity of AIPW test). *Suppose that Algorithm 1 is run on the dataset $(X_{i,n}, W_{i,n}, Y'_{i,n})_{i=1}^{n}$ while only shrinking in the X dimension so that the nth region is a subset of $\mathcal{X}^{(n)}$, the ambient space of $X^{(n)}$. Suppose that the chiseled regions satisfy Assumption D.8 and that we additionally satisfy Assumption D.19, Constraint D.9, and Condition D.10. Let $\mathcal{R}_{\tau}^{(n)}$ be the region reported by Algorithm 1 when the sample size is n, using the tests $\hat{\phi}_t^{(n)}$ given in Definition D.21. Lastly, suppose that we use the true propensities, i.e. $e_n^{(j)}(x) = e_n(x) = \mathbb{P}(W^{(n)} = 1 \mid X^{(n)} = x)$ for all j, and that the propensities satisfy strong overlap, i.e. $\eta \leq e_n(X) \leq 1 - \eta$ almost surely for some $0 < \eta \leq 0.5$. Then*

$$\limsup_{n \to \infty} \mathbb{P}\big(\mu(\mathcal{R}_{\tau}^{(n)}) \leq 0\big) \leq \alpha.$$

To reiterate: when we read Assumption D.8 in the above, we should understand it as applying to the moments of the oracle transformed outcomes, not the raw outcomes. We prove Theorem D.22 in the next subsubsection (Appendix D.5.4).

D.5.4 Proof of Theorem D.22

First, we note that the AIPW instance of chiseling which uses Definition D.21 to calculate the tests can be coupled with an oracle instance of chiseling that uses the dataset $(\tilde{X}_i, Y_i)_{i=1}^{n}$ where $\tilde{X}_i := (X_i, W_i, Y'_i)$ and which uses Definition D.5 to calculate the tests. Note that this oracle instance (1) uses $(Y_i)_{i=1}^{n}$ to calculate the test, as opposed to the AIPW instance which uses $(\hat{Y}_i)_{i=1}^{n}$, and (2) uses the true region variances $V_t := \sigma^2(\mathcal{R}_t)$ to calculate the oracle critical values C_t. We can specify these two instances of chiseling to produce the same sequence of regions $(\mathcal{R}_t)_{t=0}^{m}$ because the oracle has as much information as the AIPW instance (as encoded in $\tilde{X}_i$). We assume that these two instances also set the same sequence of target levels $(\alpha_t)_{t=0}^{m}$.

Noting that the conditions of this theorem include those necessary to apply Proposition D.11, we have that the oracle instance of chiseling is asymptotically valid. We wish to show now that the AIPW instance rejects the same region as the oracle instance with

probability going to 1. In this proof, define

$$R_t := \frac{1}{\sqrt{V_t} \cdot \sqrt{n_t}} \sum_{i : X_i \in \mathcal{R}_t} (\hat{Y}_i - Y_i), \qquad \text{(remainder)}$$

$$d_t := C_t - \sqrt{\frac{n_t}{V_t}} \cdot \mu(\mathcal{R}_t), \qquad \text{(critical value minus normalized mean)}$$

$$\hat{d}_t := \sqrt{\frac{\hat{V}_t}{V_t}} \cdot \hat{C}_t - R_t - \sqrt{\frac{n_t}{V_t}} \cdot \mu(\mathcal{R}_t), \qquad \text{(empirical variant of } d_t)$$

$$Z_t := T_t - \sqrt{\frac{n_t}{V_t}} \cdot \mu(\mathcal{R}_t) \qquad \text{(}Z\text{-stat.)}$$

(25)

We will need the following result, which is analogous to Lemma D.18.

Lemma D.23 (Uniform convergence of AIPW shifted critical values). *Under the conditions of the Theorem D.22, we have*

$$\sup_{t \in [m(n)]_0 : \alpha_t^{(n)} > 0} |d_t^{(n)} - \hat{d}_t^{(n)}| \xrightarrow{p} 0$$

where here we take the convention that the supremum over an empty set is 0.

We prove this in Appendix E.3. Now letting T_t be the oracle test statistic given by Definition D.5, note that $\hat{T}_t = \sqrt{\frac{V_t}{\hat{V}_t}} \cdot (T_t + R_t)$, and thus

$$\mathbb{1}\{T_t \leq C_t\} \neq \mathbb{1}\{\hat{T}_t \leq \hat{C}_t\}$$

$$\iff \min\left\{ C_t, \sqrt{\frac{\hat{V}_t}{V_t}} \cdot \hat{C}_t - R_t \right\} < T_t \leq \max\left\{ C_t, \sqrt{\frac{\hat{V}_t}{V_t}} \cdot \hat{C}_t - R_t \right\}$$

$$\iff \min\{d_t, \hat{d}_t\} < Z_t \leq \max\{d_t, \hat{d}_t\}.$$

Given Lemma D.23, the proof that the AIPW instance of chiseling and the oracle instance of chiseling reject the same region with probability tending to 1 follows the proof of Lemma D.12 in Appendix D.4.5 exactly. Then the asymptotic validity of the AIPW instance of chiseling follows from copying the steps of the proof of Theorem D.13 in Appendix D.4.6 exactly.

Appendix E   Proofs of supporting lemmas

E.1   Proof of Lemma D.17

For readability, we separate this into three separate proofs, one for each condition.

Proof of Lemma D.17 ($\implies$ Property D.14). Recall that Corollary 3.3 states that conditional on $\mathcal{F}_t$, the n_t pairs that fall within the region $\mathcal{R}_t$ are an i.i.d. sample of n_t pairs where each pair has the distribution $(X, Y) \mid X \in \mathcal{R}_t$. Then by the Berry–Esseen theorem, for

some absolute constant L_{BE}, the distance between the empirical CDF of the standardized sample mean in the region and the standard normal distribution can be bounded by

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P}\left(\frac{\sqrt{n_t}}{\sqrt{V_t}}(\bar{\mu}_t - \mu(\mathcal{R}_t)) \leq x \,\Big|\, \mathcal{F}_t \right) - \Phi(x) \right| \leq \frac{L_{\text{BE}} \cdot \bar{\mu}_{3,\text{cent}}(\mathcal{R}_t)}{\sigma_1^3(\mathcal{R}_t) \cdot \sqrt{n_t}} \quad \text{a.s.}$$

Then the supremum of the left-hand side over $t \in [m]_0$ such that $\alpha_t > 0$ is almost surely upper bounded by

$$\sup_{t \in [m]_0 : \alpha_t > 0} \frac{L_{\text{BE}} \cdot \bar{\mu}_{3,\text{cent}}(\mathcal{R}_t)}{\sigma_1^3(\mathcal{R}_t) \cdot \sqrt{n_t}} \leq L_{\text{BE}} \cdot N_{\min}^{-1/2} \cdot \left(\sup_{t \in [m]_0 : \alpha_t > 0} \bar{\mu}_{3,\text{cent}}(\mathcal{R}_t) \right) \left(\sup_{t \in [m]_0 : \alpha_t > 0} \sigma_1^{-3}(\mathcal{R}_t) \right)$$

where we have used the fact that $\sup_{t \in [m]_0 : \alpha_t > 0} n_t^{-1/2} = N_{\min}^{-1/2}$ by definition. Since $N_{\min} \xrightarrow{p} \infty$ by Condition D.10, it suffices that the last two supremum terms above are bounded in probability. The second supremum is directly bounded in probability by Assumption D.8. For the first supremum, we note that for all fixed $\mathcal{R} \subseteq \mathcal{X}$,

$$\bar{\mu}_{3,\text{cent}}(\mathcal{R}) = \mathbb{E}[|Y - \mu(\mathcal{R})|^3 \mid X \in \mathcal{R}] \leq 8 \cdot \mathbb{E}[|Y|^3 \mid X \in \mathcal{R}] = 8 \cdot \bar{\mu}_{3,\text{raw}}(\mathcal{R})$$

by a standard bound on centered moments by raw moments (Lemma F.12) applied to the random variable with distribution $Y \mid \{X \in \mathcal{R}\}$. Then

$$\sup_{t \in [m]_0 : \alpha_t > 0} \bar{\mu}_{3,\text{cent}}(\mathcal{R}_t) \leq 8 \cdot \sup_{t \in [m]_0 : \alpha_t > 0} \bar{\mu}_{3,\text{raw}}(\mathcal{R}_t)$$

which is bounded in probability by Assumption D.8 (note that boundedness in probability of the fourth absolute region conditional moment automatically translates to boundedness in probability for lower absolute region conditional moments; we may see this formally by applying standard L^p norm inequalities). Hence, the result follows. $\qquad\square$

Proof of Lemma D.17 ($\implies$ Property D.15). Note that by bounding the supremum of products by the product of supremums, we can almost surely bound L_{ratio} by the product of the following quantities:

$$S_1 := \sup_{s,t \in [m]_0 : \alpha_s, \alpha_t > 0} \frac{n_s}{n_t} \cdot \mathbb{1}\{p_{\text{prop}} \leq n_s/n_t \leq p_{\text{prop}}^{-1}\},$$

$$S_2 := \sup_{s,t \in [m]_0 : \alpha_s, \alpha_t > 0} V_s = \sup_{t \in [m]_0 : \alpha_t > 0} V_t,$$

$$S_3 := \sup_{s,t \in [m]_0 : \alpha_s, \alpha_t > 0} V_t^{-1} = \sup_{t \in [m]_0 : \alpha_t > 0} V_t^{-1}.$$

Note that $S_1 \leq p_{\text{prop}}^{-1}$ by definition and S_3 is bounded in probability by Assumption D.8. As for S_2, we note that $V_t \leq \bar{\mu}_{2,\text{raw}}(\mathcal{R}_t)$ almost surely. Then by Assumption D.8, which requires that $\sup_{t \in [m]_0 : \alpha_t > 0} \bar{\mu}_{4,\text{raw}}(\mathcal{R}_t)$ is bounded in probability, S_2 is also bounded in probability. Hence, the product and thus L_{ratio} are bounded in probability. $\qquad\square$

Proof of Lemma D.17 ($\implies$ Property D.16). First note that

$$\sup_{t \in [m]_0 : \alpha_t > 0} \left| \frac{\hat{V}_t}{V_t} - 1 \right| = \sup_{t \in [m]_0 : \alpha_t > 0} \left| \frac{\hat{V}_t - V_t}{V_t} \right| \leq \sup_{t \in [m]_0 : \alpha_t > 0} V_t^{-1} \cdot \sup_{t \in [m]_0 : \alpha_t > 0} |\hat{V}_t - V_t|.$$

The first supremum on the right-hand side is bounded in probability by Assumption D.8, so it suffices to show that the second supremum converges to 0 in probability. Write

$$\hat{\psi}_t = \frac{1}{n_t} \sum_{i=1}^n Y_i^2 \mathbb{1}\{X_i \in \mathcal{R}_t\}$$

and note that $\hat{V}_t = \hat{\psi}_t - \bar{\mu}_t^2$. Also note that $V_t = \bar{\mu}_{2,\text{raw}}(\mathcal{R}_t) - \mu(\mathcal{R}_t)^2$. Then

$$\sup_{t\in[m]_0:\alpha_t>0} |\hat{V}_t - V_t| \leq \sup_{t\in[m]_0:\alpha_t>0} |\hat{\psi}_t - \bar{\mu}_{2,\text{raw}}(\mathcal{R}_t)| + \sup_{t\in[m]_0:\alpha_t>0} |\bar{\mu}_t^2 - \mu(\mathcal{R}_t)^2|. \tag{26}$$

We deal with each term separately. Let $\delta > 0$ and focus on the first term. Conditional on $\mathcal{F}_t$, Corollary 3.3 tells us that $\hat{\psi}_t$ is just the simple average of n_t i.i.d. points, each of which has mean $\bar{\mu}_{2,\text{raw}}(\mathcal{R}_t)$ and variance $\sigma_2^2(\mathcal{R}_t)$. Hence $\hat{\psi}_t$ is conditionally unbiased for $\bar{\mu}_{2,\text{raw}}(\mathcal{R}_t)$ and has variance $\sigma_2^2(\mathcal{R}_t)/n_t$, so that Chebyshev's inequality yields

$$\mathbb{P}\Big(|\hat{\psi}_t - \bar{\mu}_{2,\text{raw}}(\mathcal{R}_t)| > \delta \mid \mathcal{F}_t\Big) \leq \min\left\{1, \frac{\sigma_2^2(\mathcal{R}_t)}{n_t \cdot \delta^2}\right\} \leq \min\left\{1, \frac{\bar{\mu}_{4,\text{raw}}(\mathcal{R}_t)}{n_t \cdot \delta^2}\right\}$$

and thus

$$\mathbb{P}\Big(|\hat{\psi}_t - \bar{\mu}_{2,\text{raw}}(\mathcal{R}_t)| > \delta \mid \mathcal{F}_t\Big) \cdot \mathbb{1}\{\alpha_t > 0\} \leq \min\left\{1, \frac{\bar{\mu}_{4,\text{raw}}(\mathcal{R}_t)}{N_{\min} \cdot \delta^2}\right\} \cdot \mathbb{1}\{\alpha_t > 0\}.$$

Then

$$\mathbb{P}\left(\sup_{t\in[m]_0:\alpha_t>0} |\hat{\psi}_t - \bar{\mu}_{2,\text{raw}}(\mathcal{R}_t)| > \delta\right) \leq \sum_{t=0}^m \mathbb{P}\Big(|\hat{\psi}_t - \bar{\mu}_{2,\text{raw}}(\mathcal{R}_t)| > \delta \text{ and } \alpha_t > 0\Big)$$

$$= \sum_{t=0}^m \mathbb{E}\Big[\mathbb{P}\Big(|\hat{\psi}_t - \bar{\mu}_{2,\text{raw}}(\mathcal{R}_t)| > \delta \mid \mathcal{F}_t\Big) \cdot \mathbb{1}\{\alpha_t > 0\}\Big]$$

$$\leq \mathbb{E}\left[\min\left\{1, \sup_{t\in[m]_0:\alpha_t>0} \frac{\bar{\mu}_{4,\text{raw}}(\mathcal{R}_t)}{N_{\min} \cdot \delta^2}\right\} \cdot \sum_{t=0}^m \mathbb{1}\{\alpha_t > 0\}\right]$$

$$\leq K_{\text{reg}} \cdot \mathbb{E}\left[\min\left\{1, \sup_{t\in[m]_0:\alpha_t>0} \frac{\bar{\mu}_{4,\text{raw}}(\mathcal{R}_t)}{N_{\min} \cdot \delta^2}\right\}\right]$$

$$\to 0$$

where to go to the second to last line we use Constraint D.9 to bound the sum of the indicators $\mathbb{1}\{\alpha_t > 0\}$ by K_{reg}. To go the last line we use the fact that $\sup_{t\in[m]_0:\alpha_t>0} \bar{\mu}_{4,\text{raw}}(\mathcal{R}_t)$ is bounded in probability by Assumption D.8 and that $N_{\min} \overset{p}{\to} \infty$ by Condition D.10, so that the minimum term converges to 0 in probability and hence in L^1 because it is bounded by 1.

Now we consider the second supremum in Equation (26). First, we control the convergence of the mean estimate. Again, conditional on $\mathcal{F}_t$, Corollary 3.3 tells us that $\bar{\mu}_t$ is just the simple average of n_t i.i.d. points, each of which has mean $\mu(\mathcal{R}_t)$ and variance $\sigma_1^2(\mathcal{R}_t)$, so

that $\bar{\mu}_t$ is conditionally unbiased for $\mu(\mathcal{R}_t)$ and has variance $\sigma_1^2(\mathcal{R}_t)/n_t$. Then we may carry out the same steps as before to conclude that

$$\sup_{t \in [m]_0 : \alpha_t > 0} |\bar{\mu}_t - \mu(\mathcal{R}_t)| \xrightarrow{p} 0.$$

Now since $|\bar{\mu}_t^2 - \mu(\mathcal{R}_t)^2| = |\bar{\mu}_t + \mu(\mathcal{R}_t)||\bar{\mu}_t - \mu(\mathcal{R}_t)|$, we have

$$\sup_{t \in [m]_0 : \alpha_t > 0} |\bar{\mu}_t^2 - \mu(\mathcal{R}_t)^2| \leq \sup_{t \in [m]_0 : \alpha_t > 0} |\bar{\mu}_t + \mu(\mathcal{R}_t)| \cdot \sup_{t \in [m]_0 : \alpha_t > 0} |\bar{\mu}_t - \mu(\mathcal{R}_t)|.$$

The second supremum converges to 0 in probability, so it suffices to show that the first supremum is bounded in probability. We can see this from

$$\sup_{t \in [m]_0 : \alpha_t > 0} |\bar{\mu}_t + \mu(\mathcal{R}_t)| \leq \sup_{t \in [m]_0 : \alpha_t > 0} |\bar{\mu}_t - \mu(\mathcal{R}_t)| + \sup_{t \in [m]_0 : \alpha_t > 0} 2|\mu(\mathcal{R}_t)|.$$

The first term converges to 0 in probability. Since $|\mu(\mathcal{R}_t)| \leq \bar{\mu}_{1,\text{raw}}(\mathcal{R})$, Assumption D.8 implies that the second term is bounded in probability. Thus, the sum is bounded in probability, and from this, as discussed, the second supremum in Equation (26) converges to 0 in probability. From this, the main result follows. $\qquad\square$

E.2 Proof of Lemma D.18

Before we proceed to the proof of Lemma D.18, we first establish that a few intermediate quantities converge. Let us introduce some notation to track the indices of stages where the analyst allocates non-zero error to the test. Since by Constraint 3.12 at most K_{reg} stages result in non-trivial tests, we can define t_j to be the stage number of the jth non-trivial test for $j = 1, ..., K_{\text{reg}}$. Formally,

$$t_1 := \min\{t \in [m]_0 : \alpha_t > 0\},$$
$$t_j := \min\{t \in [m]_0 : t > t_{j-1}, \alpha_t > 0\}.$$

If the set $\{t \in [m]_0 : t > t_{j-1}, \alpha_t > 0\}$ is ever empty (i.e. there are fewer than K_{reg} non-trivial tests), then we will define $t_j = t_{j-1}$. If $\alpha_t = 0$ for all t, then we will define $t_1 = ... = t_{K_{\text{reg}}} = 0$, though the choice is not important. These quantities implicitly depend on n and should formally be understood as $t_j^{(n)}$; we will suppress redundant indexing by n, for instance using the shorthand $\Delta_{t_s, t_j}^{(n)}$ while understanding that both t_s and t_j also depend on n in this expression.

Recall that ultimately in Lemma D.12 where this is applied, we assume that both the oracle and empirical versions of Algorithm 1 are run on the same dataset and make the same choices, so that $\alpha_0, ..., \alpha_n$ are identical for both versions, and thus $t_1, ..., t_{K_{\text{reg}}}$ can be referred to unambiguously for both. The next lemma tells us that it suffices to restrict our attention to these indices when calculating the truncation levels. Recall the definitions of M_t and $\hat{M}_t$ from Definitions D.5 and D.6.

Lemma E.1 (Index reduction for truncation). *For $j = 1, ..., K_{\mathrm{reg}}$,*

$$M_{t_j} = \min_{s < j : n_{t_j}/n_{t_s} \geq p_{\mathrm{prop}}} \left\{ \frac{C_{t_s} - \Delta_{t_s, t_j}}{v_{t_s, t_j}} \right\},$$

$$\hat{M}_{t_j} = \min_{s < j : n_{t_j}/n_{t_s} \geq p_{\mathrm{prop}}} \left\{ \frac{\hat{C}_{t_s} - \hat{\Delta}_{t_s, t_j}}{\hat{v}_{t_s, t_j}} \right\}.$$

Proof of Lemma E.1. We will focus on the proof of the equivalence for the oracle truncation levels. The proof for the empirical truncation levels is algebraically equivalent. If $\alpha_t = 0$ for all t, then $t_1 = ... = t_{K_{\mathrm{reg}}} = 0$, and the result follows by definition. Thus, suppose $\alpha_t > 0$ for at least one t. Note that for any $s < t$, if $\alpha_s = 0$ then

$$\frac{C_s - \Delta_{s,t}}{v_{s,t}} \geq \frac{M_s - \Delta_{s,t}}{v_{s,t}}$$

$$= \min_{u < s : n_s/n_u \geq p_{\mathrm{prop}}} \left\{ \left(\frac{C_u - \Delta_{u,s}}{v_{u,s}} - \Delta_{s,t} \right) \frac{1}{v_{s,t}} \right\}$$

$$= \min_{u < s : n_s/n_u \geq p_{\mathrm{prop}}} \left\{ \frac{C_u - (\Delta_{u,s} + v_{u,s} \Delta_{s,t})}{v_{u,s} v_{s,t}} \right\}$$

$$= \min_{u < s : n_s/n_u \geq p_{\mathrm{prop}}} \left\{ \frac{C_u - \Delta_{u,t}}{v_{u,t}} \right\}$$

where the last line follows by simple algebraic manipulations of the definitions. Now recall the definition

$$M_t = \min_{s < t : n_t/n_s \geq p_{\mathrm{prop}}} \left\{ \frac{C_s - \Delta_{s,t}}{v_{s,t}} \right\}.$$

We have just shown that for any index s where $\alpha_s = 0$, the term $(C_s - \Delta_{s,t})/v_{s,t}$ must exceed the minimum of the first $s - 1$ terms whose indices u satisfy $n_s/n_u \geq p_{\mathrm{prop}}$. Hence, the overall minimum is unchanged if we remove such indices from the minimization. That is,

$$M_t = \min_{\substack{s < t : n_t/n_s \geq p_{\mathrm{prop}} \\ \alpha_s > 0}} \left\{ \frac{C_s - \Delta_{s,t}}{v_{s,t}} \right\}.$$

But this exactly coincides with the definition of M_{t_j} when $t = t_j$ (as long as there exists at least one $\alpha_s > 0$ for $s < t$). $\qquad\square$

Next, we establish a number of intermediate convergence results. We say that a sequence of positive random variables is *positively log-bounded in probability* if both the sequence and the sequence of reciprocals are bounded in probability. We will also make use of a notion we call Φ-convergence. We elaborate on the properties of Φ-convergence in Appendix F.4. Here, we simply note that for two random sequences G_n and H_n, we say that G_n *converges to H_n under the Φ-metric* if

$$|\Phi(G_n) - \Phi(H_n)| \xrightarrow{p} 0.$$

In this case, we will write $G_n \overset{\Phi}{\leftrightarrow} H_n$. Throughout, we will make frequent use of Lemma F.25, which describes various basic operations that preserve Φ-convergence. We also recall here that by Lemma D.17, Assumption D.8, Constraint D.9, and Condition D.10 imply Properties D.14, D.15, and D.16.

Lemma E.2 (Convergence of intermediate quantities). *Let $\Delta_{s,t}$ and $v_{s,t}$ be defined as in Equation* (19) *and $\hat{\Delta}_{s,t}$ and $\hat{v}_{s,t}$ be defined analogously but with $\hat{V}_t$ used in place of V_t wherever it appears. Let $0 \leq s \leq j \leq K_{\mathrm{reg}}$ be fixed. Under Assumption D.8, Constraint D.9, and Condition D.10,*

$$\hat{\Delta}_{t_s,t_j}^{(n)} \overset{\Phi}{\leftrightarrow} \Delta_{t_s,t_j}^{(n)}.$$

Furthermore, there are positively log-bounded in probability random sequences $\hat{w}_{s,j}^{(n)}$ and $w_{s,j}^{(n)}$ such that $(\hat{w}_{s,j}^{(n)})^{-1} \overset{\Phi}{\leftrightarrow} (w_{s,j}^{(n)})^{-1}$ and

$$\hat{v}_{t_s,t_j}^{(n)} = \hat{w}_{s,j}^{(n)} \cdot \mathbb{1}\{n_{t_j}^{(n)}/n_{t_s}^{(n)} \geq p_{\mathrm{prop}}\} \quad \text{and} \quad v_{t_s,t_j}^{(n)} = w_{s,j}^{(n)} \cdot \mathbb{1}\{n_{t_j}^{(n)}/n_{t_s}^{(n)} \geq p_{\mathrm{prop}}\}.$$

Proof of Lemma E.2. We proceed in a few steps.

- First note that since $1/x$ is uniformly continuous in a neighborhood of $x = 1$, then for all $\delta > 0$, there exists an $s > 0$ such that $x \in [1 - s, 1 + s]$ implies $|1/x - 1| < \delta$. Then

$$\mathbb{P}\left(\sup_{t \in [m]_0 : \alpha_t > 0} \left| \frac{V_t}{\hat{V}_t} - 1 \right| > \delta \right) \leq \mathbb{P}\left(\exists t \in [m]_0 \text{ s.t. } \alpha_t > 0 \text{ and } \frac{\hat{V}_t}{V_t} \notin [1 - s, 1 + s] \right)$$

$$= \mathbb{P}\left(\sup_{t \in [m] : \alpha_t > 0} \left| \frac{\hat{V}_t}{V_t} - 1 \right| > s \right)$$

$$\overset{p}{\to} 0$$

by Property D.16. Then since $|\sqrt{x} - 1| \leq \sqrt{|x - 1|}$ note that for all $j = 1, ..., K_{\mathrm{reg}}$,

$$\left| \Phi\left(\sqrt{\frac{\hat{V}_{t_j}}{V_{t_j}}} \right) - \Phi(1) \right| \leq \left| \sqrt{\frac{\hat{V}_{t_j}}{V_{t_j}}} - 1 \right| \leq \sqrt{\left| \frac{\hat{V}_{t_j}}{V_{t_j}} - 1 \right|} \leq \sqrt{\sup_{t \in [m] : \alpha_t > 0} \left| \frac{\hat{V}_t}{V_t} - 1 \right|} \overset{p}{\to} 0$$

where we use the fact that Φ is 1-Lipschitz. Hence, $\sqrt{\frac{\hat{V}_{t_j}}{V_{t_j}}} \overset{\Phi}{\leftrightarrow} 1$ for all j. An analogous argument shows that $\sqrt{\frac{V_{t_j}}{\hat{V}_{t_j}}} \overset{\Phi}{\leftrightarrow} 1$ for all j.

- Note that for all $s < j$, we have that $\hat{\Delta}_{t_s,t_j} \overset{\Phi}{\leftrightarrow} \hat{\Delta}_{t_s,t_j}$ trivially, and that $\sqrt{\frac{\hat{V}_{t_s}}{V_{t_s}}} \overset{\Phi}{\leftrightarrow} 1$ as we showed above. Note also that the sequence consisting of all 1's is positively log-bounded in probability. Hence, using the multiplication rule of Lemma F.25 yields that $\sqrt{\frac{\hat{V}_{t_s}}{V_{t_s}}} \cdot \hat{\Delta}_{t_s,t_j} \overset{\Phi}{\leftrightarrow} 1 \cdot \hat{\Delta}_{t_s,t_j}$. Since $\sqrt{\frac{\hat{V}_{t_s}}{V_{t_s}}} \cdot \hat{\Delta}_{t_s,t_j} = \Delta_{t_s,t_j}$, this is equivalent to $\Delta_{t_s,t_j} \overset{\Phi}{\leftrightarrow} \hat{\Delta}_{t_s,t_j}$.

- Note that for all $s < j$, we have $\sqrt{\frac{V_{t_j}}{\hat{V}_{t_j}}} \cdot \sqrt{\frac{\hat{V}_{t_s}}{V_{t_s}}} \cdot \hat{v}_{t_s,t_j}^{-1} = v_{t_s,t_j}^{-1}$. Same as with the previous bullet, two applications of the multiplication rule of Lemma F.25 yield that $v_{t_s,t_j}^{-1} \overset{\Phi}{\leftrightarrow} \hat{v}_{t_s,t_j}^{-1}$.

97

- In the below only, for a random variable A and event E, we will appropriate the notation $A \otimes \mathbb{1}\{E\} = A\mathbb{1}\{E\} + (1 - \mathbb{1}\{E\})$. Note that $(A \otimes \mathbb{1}\{E\}) \times \mathbb{1}\{E\} = A\mathbb{1}\{E\}$. Furthermore, if $A_n \overset{\Phi}{\leftrightarrow} B_n$, then $A_n \otimes \mathbb{1}\{E_n\} \overset{\Phi}{\leftrightarrow} B_n \otimes \mathbb{1}\{E_n\}$; we can see this by first applying Lemma F.26 to conclude that $A_n\mathbb{1}\{E_n\} \overset{\Phi}{\leftrightarrow} B_n\mathbb{1}\{E_n\}$. Then since $1 - \mathbb{1}\{E_n\} \overset{\Phi}{\leftrightarrow} 1 - \mathbb{1}\{E_n\}$ trivially and is bounded in probability, the addition rule of Lemma F.25 shows that $A_n \otimes \mathbb{1}\{E_n\} \overset{\Phi}{\leftrightarrow} B_n \otimes \mathbb{1}\{E_n\}$.

- Define

$$\hat{w}_{s,j} = \hat{v}_{t_s,t_j} \otimes \mathbb{1}\{n_{t_j}/n_{t_s} \geq p_{\mathrm{prop}}\} \quad \text{and} \quad w_{s,j} = v_{t_s,t_j} \otimes \mathbb{1}\{n_{t_j}/n_{t_s} \geq p_{\mathrm{prop}}\}.$$

Then by the previous two bullets, we have $\hat{w}_{s,j}^{-1} \overset{\Phi}{\leftrightarrow} w_{s,j}^{-1}$.

- Lastly, recalling the definition of L_{ratio} from Property D.15, we have

$$v_{t_s,t_j} \cdot \mathbb{1}\{n_{t_j}/n_{t_s} \geq p_{\mathrm{prop}}\} \leq \max\{1, L_{\mathrm{ratio}}^{1/2}\},$$
$$v_{t_s,t_j}^{-1} \cdot \mathbb{1}\{n_{t_j}/n_{t_s} \geq p_{\mathrm{prop}}\} \leq \max\{1, L_{\mathrm{ratio}}^{1/2}\}.$$

This follows by the definition of L_{ratio} when $\alpha_t > 0$ for at least one t (otherwise, when $\alpha_t = 0$ for all t, then $v_{t_s,t_j} = 1$ while the right-hand sides are at least 1). Then both $w_{s,j}$ and $w_{s,j}^{-1}$ are bounded by $\max\{1, L_{\mathrm{ratio}}^{1/2}\}$ which is bounded in probability by Property D.15. Thus $w_{s,j}$ is positively log-bounded in probability. By the transitivity of positively log-boundedness in probability (Lemma F.18), we also have that $\hat{w}_{s,j}$ is positively log-bounded in probability.

$\square$

Now we may use these results to establish the uniform convergence of the critical values.

Lemma E.3 (Uniform convergence of critical values). *Under Assumption D.8, Constraint D.9, and Condition D.10, we have*

$$\sup_{t \in [m(n)]_0 : \alpha_t^{(n)} > 0} |\hat{C}_t^{(n)} - C_t^{(n)}| \overset{p}{\to} 0$$

where here we take the convention that the supremum over an empty set is 0.

Proof of Lemma E.3. First define the indicator $H := \mathbb{1}\{\exists t \text{ s.t. } \alpha_t > 0\}$, and define $C_t' := C_t \cdot H$ and $\hat{C}_t' := \hat{C}_t \cdot H$. Here we take the convention that $\infty \cdot 0 = 0$. Then

$$\sup_{t \in [m]_0 : \alpha_t > 0} |\hat{C}_t - C_t| = \sup_{j = 1, \dots, K_{\mathrm{reg}}} |\hat{C}_{t_j}' - C_{t_j}'|$$

since if $H = 1$ then the supremum on the left-hand side is precisely over the indices $t_1, \dots, t_{K_{\mathrm{reg}}}$ (with possibly some redundancy), and otherwise both sides of the above are 0.

Next note that for $j = 1, \dots, K_{\mathrm{reg}}$ and for all y,

$$\Phi^{-1}(1 - \alpha_{t_j}; y) \cdot H \leq \Phi^{-1}(1 - \alpha_{t_j}) \cdot H \leq \Phi^{-1}(1 - \alpha_{\min}) < \infty.$$

98

The second inequality follows since if $H = 1$, then $\alpha_{t_j} > 0$ and hence $\alpha_{t_j} \geq \alpha_{\min}$ by Constraint D.9. Hence $C'_{t_j}, \hat{C}'_{t_j} \in [0, \Phi^{-1}(1 - \alpha_{\min})]$ a.s., and are hence also bounded in probability.

Next, we will show by induction that $\hat{C}'_{t_j} \overset{\Phi}{\leftrightarrow} C'_{t_j}$ for $j = 1, ..., K_{\text{reg}}$. The base case of $j = 1$ follows from the fact that $\hat{M}_{t_1} = M_{t_1} = \infty$ by Lemma E.1, and so $\hat{C}_{t_1} = C_{t_1}$, which also implies that $\hat{C}'_{t_1} = C'_{t_1}$. Now suppose $\hat{C}'_{t_s} = C'_{t_s}$ for all $s < j$. We will show that $\hat{C}'_{t_j} \overset{\Phi}{\leftrightarrow} C'_{t_j}$.

Let $w_{s,j}$ and $\hat{w}_{s,j}$ be defined as in Lemma E.2. Since C'_{t_s} is bounded in probability, the subtraction rule of Lemma F.25, the inductive hypothesis, and Lemma E.2 yield that $\hat{C}'_{t_s} - \hat{\Delta}_{t_s,t_j} \overset{\Phi}{\leftrightarrow} C'_{t_s} - \Delta_{t_s,t_j}$ for every $s < j$. Then this fact, the multiplication rule of Lemma F.25, and Lemma E.2 yield that $\frac{\hat{C}'_{t_s} - \hat{\Delta}_{t_s,t_j}}{\hat{w}_{s,j}} \overset{\Phi}{\leftrightarrow} \frac{C'_{t_s} - \Delta_{t_s,t_j}}{w_{s,j}}$ for each $s < j$. Then by repeated application of the minimum rule of Lemma F.25,

$$\hat{m}_j := \min_{s<j:n_{t_j}/n_{t_s} \geq p_{\text{prop}}} \left\{ \frac{\hat{C}'_{t_s} - \hat{\Delta}_{t_s,t_j}}{\hat{w}_{s,j}} \right\} \overset{\Phi}{\leftrightarrow} \min_{s<j:n_{t_j}/n_{t_s} \geq p_{\text{prop}}} \left\{ \frac{C'_{t_s} - \Delta_{t_s,t_j}}{w_{s,j}} \right\} =: m_j.$$

Also note that in the above, we may replace $w_{s,j}$ with v_{t_s,t_j} and $\hat{w}_{s,j}$ with $\hat{v}_{t_s,t_j}$, since these quantities agree whenever $n_{t_j}/n_{t_s} \geq p_{\text{prop}}$.

Now letting $\text{Clip}_{a,b}(x) = \max\{a, \min\{x, b\}\}$, using the clipped truncated normal quantile rule of Lemma F.25 and the fact that $\alpha_{t_j} \overset{\Phi}{\leftrightarrow} \alpha_{t_j}$ trivially, we have

$$\text{Clip}_{0,\infty}(\Phi^{-1}(\text{Clip}_{0,b}(1 - \alpha_{t_j})) \cdot \Phi(\hat{m}_j)) \overset{\Phi}{\leftrightarrow} \text{Clip}_{0,\infty}(\Phi^{-1}(\text{Clip}_{0,b}(1 - \alpha_{t_j})) \cdot \Phi(m_j))$$

where $b = 1 - \alpha_{\min}$. Observe that

$$\hat{C}'_{t_j} = \max\{0, \Phi^{-1}(1 - \alpha_{t_j}; \hat{M}_{t_j})\} \cdot H = \text{Clip}_{0,\infty}(\Phi^{-1}(\text{Clip}_{0,b}(1 - \alpha_{t_j})) \cdot \Phi(\hat{m}_j) \cdot H.$$

The first equality follows by definition. For the second equality, note that if $H = 0$, then both sides of the above are 0. If $H = 1$, then $\alpha_{t_j} > 0$ and hence $\alpha_{t_j} \geq \alpha_{\min}$ almost surely, and so $1 - \alpha_{t_j} = \text{Clip}_{0,b}(1 - \alpha_{t_j})$. Also, if $H = 1$, then $\hat{C}'_{t_s} = \hat{C}_{t_s}$ for all $s < j$, and hence $\hat{m}_j = \hat{M}_{t_j}$ by definition of the former and Lemma E.1. So the second equality follows after making the appropriate substitutions. By the same reasoning,

$$C'_{t_j} = \max\{0, \Phi^{-1}(1 - \alpha_{t_j}; M_{t_j})\} \cdot H = \text{Clip}_{0,\infty}(\Phi^{-1}(\text{Clip}_{0,b}(1 - \alpha_{t_j})) \cdot \Phi(m_j) \cdot H.$$

Finally, Lemma F.26 implies that $\hat{C}'_{t_j} \overset{\Phi}{\leftrightarrow} C'_{t_j}$, which completes the proof of the inductive step.

Hence, we have shown that $\hat{C}'_{t_j} \overset{\Phi}{\leftrightarrow} C'_{t_j}$ for $j = 1, ..., K_{\text{reg}}$. Recalling that $\hat{C}'_{t_j}$ and C'_{t_j} both lie in the bounded interval $[0, \Phi^{-1}(1 - \alpha_{\min})]$ a.s., we may use the fact that bounded Φ-convergent sequences converge in the ordinary sense (Lemma F.27), so that $|\hat{C}'_{t_j} - C'_{t_j}| \overset{p}{\to} 0$. Finally, we have

$$\sup_{t \in [m]_0 : \alpha_t > 0} |\hat{C}_t - C_t| = \sup_{j=1,...,K_{\text{reg}}} |\hat{C}'_{t_j} - C'_{t_j}| \overset{p}{\to} 0$$

where the final convergence follows from the continuous mapping theorem applied to the function $g(x_1, ..., x_{K_{\text{reg}}}) = \sup_{j=1,...,K_{\text{reg}}} x_j$. □

We are finally ready to prove Lemma D.18.

Proof of Lemma D.18. Recall, as in the proof of Lemma E.3, that by Constraint D.9, whenever $\alpha_t > 0$, we also have $\alpha_t \geq \alpha_{\min}$. So for $j = 1, ..., K_{\text{reg}}$,

$$\Phi^{-1}(1 - \alpha_{t_j}; y) \leq \Phi^{-1}(1 - \alpha_{t_j}) \leq \Phi^{-1}(1 - \alpha_{\min}) < \infty.$$

Hence $C_{t_j}, \hat{C}_{t_j} \in [0, \Phi^{-1}(1 - \alpha_{\min})]$ a.s. Then

$$
\begin{aligned}
&\sup_{t \in [m]_0 : \alpha_t > 0} |d_t - \hat{d}_t| \\
&= \sup_{t \in [m]_0 : \alpha_t > 0} \left| C_t - \sqrt{\frac{\hat{V}_t}{V_t}} \cdot \hat{C}_t \right| \\
&\leq \sup_{t \in [m]_0 : \alpha_t > 0} \left| C_t - \sqrt{\frac{\hat{V}_t}{V_t}} \cdot C_t \right| + \left| \sqrt{\frac{\hat{V}_t}{V_t}} \cdot C_t - \sqrt{\frac{\hat{V}_t}{V_t}} \cdot \hat{C}_t \right| \\
&\leq \Phi^{-1}(1 - \alpha_{\min}) \sup_{t \in [m]_0 : \alpha_t > 0} \left| \sqrt{\frac{\hat{V}_t}{V_t}} - 1 \right| + \sup_{t \in [m]_0 : \alpha_t > 0} \left| \sqrt{\frac{\hat{V}_t}{V_t}} \right| \sup_{t \in [m]_0 : \alpha_t > 0} |C_t - \hat{C}_t|.
\end{aligned}
\tag{27}
$$

Then since $|\sqrt{x} - 1| \leq \sqrt{|x - 1|}$,

$$
\sup_{t \in [m]_0 : \alpha_t > 0} \left| \sqrt{\frac{\hat{V}_t}{V_t}} - 1 \right| \leq \sup_{t \in [m]_0 : \alpha_t > 0} \sqrt{\left| \frac{\hat{V}_t}{V_t} - 1 \right|} = \sqrt{\sup_{t \in [m]_0 : \alpha_t > 0} \left| \frac{\hat{V}_t}{V_t} - 1 \right|} \xrightarrow{p} 0.
$$

by the continuous mapping theorem and Property D.16. Next, note that

$$
\sup_{t \in [m]_0 : \alpha_t > 0} \left| \sqrt{\frac{\hat{V}_t}{V_t}} \right| \leq 1 + \sup_{t \in [m]_0 : \alpha_t > 0} \left| \sqrt{\frac{\hat{V}_t}{V_t}} - 1 \right| \xrightarrow{p} 1.
$$

Lastly, by Lemma E.3 we have $\sup_{t \in [m]_0 : \alpha_t > 0} |C_t - \hat{C}_t| \xrightarrow{p} 0$. Then applying these results to the last line in Equation (27) along with Slutsky's theorem concludes the lemma. $\square$

E.3   Proof of Lemma D.23

The proof ideas here are very similar to the proof of Lemma D.18 in Appendix E.2. In fact, we borrow the definition of t_j for $j = 1, ..., K_{\text{reg}}$ at the beginning of Appendix E.2, and we may utilize Lemma E.1 here also since it is an algebraic fact. Furthermore, we will also require the notion of *positively log-bounded in probability* and Φ-convergence. To avoid repeating too much, we refer the reader to the part of Appendix E.2 preceding the statement of Lemma E.2. Fundamentally, the proof framework is the same, except that there are more intermediate quantities to address.

Lemma E.4 (Convergence of AIPW remainders). *Recall the definition of R_t from Equation (25), and further define*

$$J_t := \frac{1}{n_t} \sum_{i:X_i \in \mathcal{R}_t} (\hat{Y}_i^2 - Y_i^2).$$

Then under the conditions of Theorem D.22, both $\sup_{t \in [m]_0:\alpha_t > 0} |R_t|$ and $\sup_{t \in [m]_0:\alpha_t > 0} |J_t|$ converge to 0 in probability. Furthermore, $\sup_{t \in [m]_0:\alpha_t > 0} |\hat{\mu}_t - \bar{\mu}_t| \overset{p}{\to} 0$.

Proof of Lemma E.4. We first define

$$\tilde{R}_t := \frac{1}{\sqrt{n_t}} \sum_{i:X_i \in \mathcal{R}_t} (\hat{Y}_i - Y_i) = \frac{1}{\sqrt{n_t}} \sum_{j=1}^{K_{\text{cv}}} \sum_{i:X_i \in \mathcal{R}_t} (\hat{Y}_{i,1} - Y_{i,1} - \hat{Y}_{i,0} + Y_{i,0}) \mathbb{1}\{i \in \mathcal{I}_j\}$$

and note that since $|R_t| = V_t^{-1/2} \times |\tilde{R}_t|$,

$$\sup_{t:\alpha_t > 0} |R_t| \leq \sqrt{\sup_{t:\alpha_t > 0} V_t^{-1}} \times \sup_{t:\alpha_t > 0} |\tilde{R}_t|.$$

The first supremum on the right-hand side is bounded in probability by Assumption D.8, so it suffices to show that the second supremum on the right-hand side converges to 0. In fact, by the triangle inequality, it suffices to show that

$$\sup_{t:\alpha_t > 0} |\tilde{R}_t^{(j,w)}| \overset{p}{\to} 0 \quad \text{where} \quad \tilde{R}_t^{(j,w)} := \frac{1}{\sqrt{n_t}} \sum_{i \in \mathcal{I}_j : X_i \in \mathcal{R}_t} (\hat{Y}_{i,w} - Y_{i,w}).$$

Without loss of generality, we will let $j = 1$ and $w = 1$. The proof when $w = 0$ is essentially identical. First note that if $i \in \mathcal{I}_1$ then we can expand $\hat{Y}_{i,1} - Y_{i,1}$ as

$$\hat{g}_1^{(1)}(X_i) + \frac{W_i(Y_i' - \hat{g}_1^{(1)}(X_i))}{e(X_i)} - g_1(X_i) - \frac{W_i(Y_i' - g(X_i))}{e(X_i)} = (\hat{g}_1^{(1)}(X_i) - g_1(X_i))\left(1 - \frac{W_i}{e(X_i)}\right).$$

Define $\mathcal{G}_t := \sigma(\mathcal{F}_t, \mathcal{I}_1, (X_i, W_i, Y_i')_{i \notin \mathcal{I}_1}, (X_i)_{i=1}^n)$. Conditionally on $\mathcal{G}_t$, we have that $\hat{g}_1^{(1)}(X_i) - g_1(X_i)$ is fixed for all $i \in \mathcal{I}_1$. Furthermore, we still have $\mathbb{P}(W_i = 1 \mid \mathcal{G}) = e(X_i)$ for all $i \in \mathcal{I}_1$ such that $X_i \in \mathcal{R}_t$ and that $(W_i)_{i \in \mathcal{I}_1:X_i \in \mathcal{R}_t}$ are independent. This is because conditioning on $\mathcal{F}_t$ and $\mathcal{I}_1$ tells us that $(X_i, W_i, Y_i')_{i \in \mathcal{I}_1:X_i \in \mathcal{R}_t}$ is an i.i.d. sample from $(X, W, Y') \mid X \in \mathcal{R}_t$ by Corollary 3.3 and the fact that the folds are generated independently. Further conditioning on the data in $\mathcal{I}_1^c$ does not affect the distribution of $(X_i, W_i, Y_i')_{i \in \mathcal{I}_1:X_i \in \mathcal{R}_t}$ since $\mathcal{I}_1$ and $\mathcal{I}_1^c$ are disjoint, and further conditioning on $(X_i)_{i=1}^n$ fixes the values of X_i but does not change the conditional distribution of $W_i \mid X_i$ for $i \in \mathcal{I}_1$ such that $X_i \in \mathcal{R}_t$. Then $1 - W_i/e(X_i)$ has mean 0 conditionally on $\mathcal{G}_t$ and

$$\mathbb{E}\left[\tilde{R}_t^{(1,1)} \,\middle|\, \mathcal{G}_t\right] = 0$$

101

and

$$\beta_t := \text{Var}(\tilde{R}_t^{(1,1)} \mid \mathcal{G}_t) = \frac{1}{n_t} \sum_{i \in \mathcal{I}_1 : X_i \in \mathcal{R}_t} \text{Var}(\hat{Y}_{i,w} - Y_{i,w} \mid \mathcal{G}_t)$$

$$= \frac{1}{n_t} \sum_{i \in \mathcal{I}_1 : X_i \in \mathcal{R}_t} (\hat{g}_1^{(1)}(X_i) - g_1(X_i))^2 \cdot \mathbb{E}\left[\left(1 - \frac{W_i}{e(X_i)}\right)^2 \Bigg| \mathcal{G}_t\right]$$

$$= \frac{1}{n_t} \sum_{i \in \mathcal{I}_1 : X_i \in \mathcal{R}_t} (\hat{g}_1^{(1)}(X_i) - g_1(X_i))^2 \cdot \frac{1 - e(X_i)}{e(X_i)}$$

$$\leq \frac{|\{i \in \mathcal{I}_1 : X_i \in \mathcal{R}_t\}|}{n_t} \cdot L_{\text{stable}}^2 \cdot \frac{1 - \eta}{\eta}$$

$$\leq L_{\text{stable}}^2 \cdot \frac{1 - \eta}{\eta}$$

where we recall the definition of L_{stable} from Assumption D.19. This is true for all $t = 0, ..., m$ as long as $\alpha_t > 0$, so $\beta^+ := \sup_{t \in [m]_0 : \alpha_t > 0} \beta_t \leq L_{\text{stable}}^2 \cdot \frac{1 - \eta}{\eta}$. Since the latter converges to 0 in probability, for any $\delta, \epsilon > 0$ let $b = (\epsilon \delta^2)/(2K_{\text{reg}})$ and choose N so that $\mathbb{P}(\beta^+ > b) \leq \epsilon/2$ for all $n \geq N$. Define the event $E := \{\beta^+ \leq b\}$. Then

$$\mathbb{P}\left(\sup_{t : \alpha_t > 0} |\tilde{R}_t^{(1,1)}| > \delta, E\right) \leq \sum_{t=0}^{m} \mathbb{P}(|\tilde{R}_t^{(1,1)}| > \delta, \alpha_t > 0, E)$$

$$= \sum_{t=0}^{m} \mathbb{E}[\mathbb{1}\{\alpha_t > 0\}\mathbb{P}(|\tilde{R}_t^{(1,1)}| > \delta, E \mid \mathcal{G}_t)]$$

$$\leq \sum_{t=0}^{m} \mathbb{E}[\mathbb{1}\{\alpha_t > 0\}\mathbb{P}(|\tilde{R}_t^{(1,1)}| > \delta, \beta_t \leq b \mid \mathcal{G}_t)]$$

$$\leq \sum_{t=0}^{m} \mathbb{E}\left[\mathbb{1}\{\alpha_t > 0\} \cdot \frac{b}{\delta^2}\right]$$

$$\leq \frac{b}{\delta^2} \cdot K_{\text{reg}}$$

$$= \epsilon/2.$$

To go to the fourth line we have noted that $\mathbb{P}(|\tilde{R}_t^{(1,1)}| > \delta, \beta_t \leq b \mid \mathcal{G}_t) = \mathbb{P}(|\tilde{R}_t^{(1,1)}| > \delta \mid \mathcal{G}_t)\mathbb{1}\{\beta_t \leq b\}$ since β_t is $\mathcal{G}_t$-measurable and then applied the conditional Chebyshev inequality. Then

$$\mathbb{P}\left(\sup_{t : \alpha_t > 0} |\tilde{R}_t^{(1,1)}| > \delta\right) \leq \mathbb{P}\left(\sup_{t : \alpha_t > 0} |\tilde{R}_t^{(1,1)}| > \delta, E\right) + \mathbb{P}(E^{\mathsf{c}}) \leq \epsilon/2 + \epsilon/2 = \epsilon$$

for all $n \geq N$ as desired. Then again, repeating the above across all j and w and bounding using the triangle inequality, we obtain that $\sup_{t : \alpha_t > 0} |\tilde{R}_t| \xrightarrow{p} 0$. Note that as an immediate consequence we obtain

$$\sup_{t : \alpha_t > 0} |\hat{\mu}_t - \bar{\mu}_t| \leq \sup_{t : \alpha_t > 0} |\tilde{R}_t| \xrightarrow{p} 0.$$

Lastly, we turn our attention to $\sup_{t:\alpha_t>0}|J_t|$. Note that by Cauchy-Schwarz we have

$$|J_t| = \frac{1}{n_t}\left|\sum_{i:X_i\in\mathcal{R}_t}(\hat{Y}_i+Y_i)(\hat{Y}_i-Y_i)\right| \leq \sqrt{\frac{1}{n_t}\sum_{i:X_i\in\mathcal{R}_t}(\hat{Y}_i+Y_i)^2}\sqrt{\frac{1}{n_t}\sum_{i:X_i\in\mathcal{R}_t}(\hat{Y}_i-Y_i)^2}$$

and thus

$$\sup_{t:\alpha_t>0}|J_t| \leq \sqrt{\sup_{t:\alpha_t>0}\frac{1}{n_t}\sum_{i:X_i\in\mathcal{R}_t}(\hat{Y}_i+Y_i)^2}\sqrt{\sup_{t:\alpha_t>0}\frac{1}{n_t}\sum_{i:X_i\in\mathcal{R}_t}(\hat{Y}_i-Y_i)^2}.$$

We can write $\hat{Y}_i + Y_i = 2Y_i + (\hat{Y}_i - Y_i)$ and use the triangle inequality for norms to obtain

$$\sqrt{\sup_{t:\alpha_t>0}\frac{1}{n_t}\sum_{i:X_i\in\mathcal{R}_t}(\hat{Y}_i+Y_i)^2} \leq 2\sqrt{\sup_{t:\alpha_t>0}\frac{1}{n_t}\sum_{i:X_i\in\mathcal{R}_t}Y_i^2} + \sqrt{\sup_{t:\alpha_t>0}\frac{1}{n_t}\sum_{i:X_i\in\mathcal{R}_t}(\hat{Y}_i-Y_i)^2}.$$

It is straightforward to show that the first supremum term on the right-hand side is bounded in probability. We can show this by again expanding $Y_i = \mu(\mathcal{R}_t) + (Y_i - \mu(\mathcal{R}_t))$ and applying the triangle inequality for norms, then using the probability bounds on the supremum over moments granted by Assumption D.8. The second supremum term is repeated from the previous display, and thus to show that $\sup_{t:\alpha_t>0}|J_t| \overset{p}{\to}$ it suffices to show that it converges to 0 in probability as then $\sup_{t:\alpha_t>0}|J_t|$ is bounded by the product of a term that is bounded in probability and a term that converges to 0 in probability. We focus on this term now. We expand $(\hat{Y}_i - Y_i)^2 = (\hat{Y}_{i,1} - Y_{i,1})^2 + 2(\hat{Y}_{i,1} - Y_{i,1})(\hat{Y}_{i,0} - Y_{i,0}) + (\hat{Y}_{i,0} - Y_{i,0})^2$ to bound the supremum by the sum of the following three terms that we now characterize.

$$\sup_{t:\alpha_t>0}\frac{1}{n_t}\sum_{i:X_i\in\mathcal{R}_t}(\hat{Y}_{i,1}-Y_{i,1})^2 \leq L_{\text{stable}}^2 \cdot \sup_{t:\alpha_t>0}\frac{1}{n_t}\sum_{i:X_i\in\mathcal{R}_t}\left(1-\frac{W_i}{e(X_i)}\right)^2$$

$$\leq L_{\text{stable}}^2 \cdot \sup_{t:\alpha_t>0}\frac{1}{n_t}\sum_{i:X_i\in\mathcal{R}_t}\left(\frac{1-\eta}{\eta}\right)^2$$

$$\leq L_{\text{stable}}^2 \cdot \left(\frac{1-\eta}{\eta}\right)^2$$

$$\overset{p}{\to} 0.$$

A similar calculation shows that

$$\sup_{t:\alpha_t>0}\frac{1}{n_t}\sum_{i:X_i\in\mathcal{R}_t}(\hat{Y}_{i,0}-Y_{i,0})^2 \leq L_{\text{stable}}^2 \cdot \left(\frac{1-\eta}{\eta}\right)^2 \overset{p}{\to} 0.$$

Lastly, note that by Cauchy-Schwarz,

$$\sup_{t:\alpha_t>0} \frac{2}{n_t} \left| \sum_{i:X_i\in\mathcal{R}_t} (\hat{Y}_{i,1} - Y_{i,1})(\hat{Y}_{i,0} - Y_{i,0}) \right|$$

$$\leq 2 \sup_{t:\alpha_t>0} \sqrt{\frac{1}{n_t} \sum_{i:X_i\in\mathcal{R}_t} (\hat{Y}_{i,1} - Y_{i,1})^2} \sqrt{\frac{1}{n_t} \sum_{i:X_i\in\mathcal{R}_t} (\hat{Y}_{i,0} - Y_{i,0})^2}$$

$$\leq 2 \sqrt{\sup_{t:\alpha_t>0} \frac{1}{n_t} \sum_{i:X_i\in\mathcal{R}_t} (\hat{Y}_{i,1} - Y_{i,1})^2} \sqrt{\sup_{t:\alpha_t>0} \frac{1}{n_t} \sum_{i:X_i\in\mathcal{R}_t} (\hat{Y}_{i,0} - Y_{i,0})^2}$$

$$\xrightarrow{p} 0$$

by the bounds we just showed. Combining these bounds yields $\sup_{t:\alpha_t>0}|J_t| \xrightarrow{p} 0$. $\qquad\square$

Lemma E.5 (Convergence of AIPW variances). *Recall that $V_t := \sigma^2(\mathcal{R}_t)$ and $\hat{V}_t$ is defined as in Equation* (24). *Then under the conditions of Theorem* D.22,

$$\sup_{t\in[m]_0:\alpha_t>0} \left| \frac{\hat{V}_t}{V_t} - 1 \right| \xrightarrow{p} 0.$$

Proof of Lemma E.5. First note that as in the proof of Property D.16 in Appendix E.1, it suffices to show that

$$\sup_{t\in[m]_0:\alpha_t>0} |\hat{V}_t - V_t| \xrightarrow{p} 0.$$

Letting

$$\tilde{V}_t := \frac{1}{n_t} \sum_{i:X_i\in\mathcal{R}_t} (Y_i - \bar{\mu}_t)^2$$

note that

$$\sup_{t\in[m]_0:\alpha_t>0} |\hat{V}_t - V_t| \leq \sup_{t\in[m]_0:\alpha_t>0} |\hat{V}_t - \tilde{V}_t| + \sup_{t\in[m]_0:\alpha_t>0} |\tilde{V}_t - V_t|$$

The second supremum term on the right-hand side is exactly that which is addressed in the proof of Property D.16 in Appendix E.1, where we show that it converges to 0 in probability. Thus, we focus on the first term. Note that we may write

$$\hat{V}_t = \left(\frac{1}{n_t} \sum_{i:X_i\in\mathcal{R}_t} \hat{Y}_i^2 \right) - \hat{\mu}_t^2 \quad \text{and} \quad \tilde{V}_t = \left(\frac{1}{n_t} \sum_{i:X_i\in\mathcal{R}_t} Y_i^2 \right) - \bar{\mu}_t^2.$$

Then we may bound

$$\sup_{t:\alpha_t>0} |\hat{V}_t - \tilde{V}_t| \leq \sup_{t:\alpha_t>0} \left| \frac{1}{n_t} \sum_{i:X_i\in\mathcal{R}_t} (\hat{Y}_i^2 - Y_i^2) \right| + \sup_{t:\alpha_t>0} |\hat{\mu}_t^2 - \bar{\mu}_t^2|.$$

The first supremum on the right-hand side converges to 0 in probability by Lemma E.4. For the second term, we may write

$$\sup_{t:\alpha_t>0} |\hat{\mu}_t^2 - \bar{\mu}_t^2| \leq \sup_{t:\alpha_t>0} |\hat{\mu}_t + \bar{\mu}_t| \times \sup_{t:\alpha_t>0} |\hat{\mu}_t - \bar{\mu}_t|.$$

The first supremum is bounded in probability since it can be coupled to $2\sup_{t:\alpha_t>0}|\mu(\mathcal{R}_t)| \le 2\sup_{t:\alpha_t>0}\bar{\mu}_{1,\mathrm{raw}}(\mathcal{R}_t)$ which is bounded in probability by Assumption D.8. The second converges to 0 by Lemma E.4. □

We now state and prove the analogue of Lemma E.2 adapted for our current purposes.

Lemma E.6 (Convergence of AIPW intermediate quantities). *Let $\Delta_{s,t}$ and $v_{s,t}$ be defined as in Equation* (19) *and $\hat{\Delta}_{s,t}$ and $\hat{v}_{s,t}$ be defined as in Equation* (24). *Let $0 \le s \le j \le K_{\mathrm{reg}}$ be fixed. Under the conditions of Theorem D.22,*

$$\hat{\Delta}^{(n)}_{t_s,t_j} \overset{\Phi}{\leftrightarrow} \Delta^{(n)}_{t_s,t_j}.$$

Furthermore, there are positively log-bounded in probability random sequences $\hat{w}^{(n)}_{s,j}$ and $w^{(n)}_{s,j}$ such that $(\hat{w}^{(n)}_{s,j})^{-1} \overset{\Phi}{\leftrightarrow} (w^{(n)}_{s,j})^{-1}$ and

$$\hat{v}^{(n)}_{t_s,t_j} = \hat{w}^{(n)}_{s,j} \cdot \mathbb{1}\{n^{(n)}_{t_j}/n^{(n)}_{t_s} \ge p_{\mathrm{prop}}\} \quad \text{and} \quad v^{(n)}_{t_s,t_j} = w^{(n)}_{s,j} \cdot \mathbb{1}\{n^{(n)}_{t_j}/n^{(n)}_{t_s} \ge p_{\mathrm{prop}}\}.$$

Proof of Lemma E.6. First note that the proofs of all of the properties of $\hat{V}_{t_s}$, V_{t_s}, their ratios, $\hat{v}_{t_s,t_j}$, and v_{t_s,t_j} can be exactly lifted from the proof of Lemma E.2. Also, the construction of $\hat{w}^{-1}_{s,j}$ and $w^{-1}_{s,j}$ can be exactly copied from the proof of Lemma E.2, and the proof of their properties applied. Thus, we focus on the first part of the theorem involving $\hat{\Delta}_{t_s,t_j}$ and Δ_{t_s,t_j}. Define

$$A_{s,t} := \frac{1}{\sqrt{n_s}} \sum_{i:X_i \in \mathcal{R}_s \setminus \mathcal{R}_t} (\hat{Y}_i - Y_i)$$

noting that $A_{s,t}$ depends on n but we have suppressed dependence on n. We have by the multiplication rule of Lemma F.25 that

$$\sqrt{\frac{\hat{V}_{t_s}}{V_{t_s}}} \cdot \hat{\Delta}_{t_s,t_j} \overset{\Phi}{\leftrightarrow} \hat{\Delta}_{t_s,t_j}$$

since $\sqrt{\frac{\hat{V}_{t_s}}{V_{t_s}}} \overset{\Phi}{\leftrightarrow} 1$ (as shown in the proof of Lemma E.2) which is log bounded in probability. In a moment we will show that $V^{-1/2}_{t_s} A_{t_s,t_j} \overset{p}{\to} 0$. Then we argue that this suffices, since

$$\left|\Phi(\Delta_{t_s,t_j}) - \Phi(\hat{\Delta}_{t_s,t_j})\right| = \left|\Phi\left(\sqrt{\frac{\hat{V}_{t_s}}{V_{t_s}}} \cdot \hat{\Delta}_{t_s,t_j} - \frac{1}{\sqrt{V_{t_s}}} \cdot A_{t_s,t_j}\right) - \Phi(\hat{\Delta}_{t_s,t_j})\right|$$

$$\le \left|\Phi\left(\sqrt{\frac{\hat{V}_{t_s}}{V_{t_s}}} \cdot \hat{\Delta}_{t_s,t_j}\right) - \Phi(\hat{\Delta}_{t_s,t_j})\right| + |V^{-1/2}_{t_s} A_{t_s,t_j}|$$

by the fact the Φ is 1-Lipschitz. Note that the first absolute value term in the last expression converges to 0 by the definition of Φ-convergence, and the second term converges in probability to 0 as we will show. Then, $\Delta_{t_s,t_j} \overset{\Phi}{\leftrightarrow} \hat{\Delta}_{t_s,t_j}$, as desired.

Note that

$$|V_{t_s}^{-1/2} A_{t_s,t_j}| \leq \sqrt{\sup_{t:\alpha_t>0} V_t^{-1}} \cdot A_{t_s,t_j}$$

and the supremum is bounded in probability by Assumption D.8 (this is true even when $\alpha_t = 0$ for all t since then $A_{t_s,t_j} = 0$ by our convention). So it suffices to show that $A_{t_s,t_j} \xrightarrow{p} 0$. Let us break this term down even further:

$$\tilde{A}_{s,t}^{(k,w)} := \frac{1}{\sqrt{n_s}} \sum_{i \in \mathcal{I}_k : X_i \in \mathcal{R}_s \backslash \mathcal{R}_t} (\hat{Y}_{i,w} - Y_{i,w})$$

and note that $A_{s,t}$ is a fixed linear combination of the above over $w = 0, 1$ and $k = 1, ..., K_{\mathrm{cv}}$. So it suffices to show that $\tilde{A}_{t_s,t_j}^{(k,w)} \xrightarrow{p} 0$ for all such w and k. Without loss of generality, we will consider $w = 1$ and $k = 1$ and drop the superscript on $(k, w) = (1, 1)$ for notational clarity.

Next, we slightly redefine t_j for $j = 1, ..., K_{\mathrm{reg}}$ and introduce some conventions to make them more convenient to work with. Let

$$t_1 := \min\{t \in [m]_0 : \alpha_t > 0\},$$
$$t_j := \min\{t \in [m]_0 : t > t_{j-1}, \alpha_t > 0\},$$

but where now we take the convention that the minimum of an empty set is $m + 1$. Then take the convention that $\tilde{A}_{s,t} = 0$ if either s or t is equal to $m + 1$. This does not modify the value of $\tilde{A}_{t_s,t_j}$ from the definition we had before, since t_s or t_j equals $m + 1$ if and only if $t_s = t_j$ which implies that $\tilde{A}_{t_s,t_j} = 0$, as we can check by cases. This definition ensures that for all $j = 1, ..., K_{\mathrm{reg}}$, t_j is a stopping time with respect to $(\mathcal{F}_t)_{t=0}^m$.

Define $\mathcal{G}_t := \sigma(\mathcal{F}_t, \mathcal{I}_1, (X_i, W_i, Y_i')_{i \notin \mathcal{I}_1}, (X_i)_{i=1}^n)$ for $t = 0, ..., m + 1$ where we let $\mathcal{F}_{m+1} := \mathcal{F}_m$. Subsequently, we will work conditionally on $\mathcal{G}_{t_s}$, so treat all $\mathcal{G}_{t_s}$-measurable quantities as fixed. In particular, we may treat t_s as a fixed value. Let us slightly simplify notation again and define the following process $r = t_s, ..., m$:

$$M_r := \tilde{A}_{t_s,r}$$

and $M_{m+1} := M_m$. First note that $|\tilde{A}_{t_s,t_j}| \leq |M_{t_j}|$ almost surely (with inequality only possible if $t_j = m + 1$), so it suffices to show that the latter converges to 0 in probability. Our idea is to show that $\mathbb{E}[M_{t_j} \mid \mathcal{G}_{t_s}] = 0$ and $\mathrm{Var}(M_{t_j} \mid \mathcal{G}_{t_s}) \xrightarrow{p} 0$. Then the result follows by Lemma F.13. We first show that conditionally on $\mathcal{G}_{t_s}$, $(M_r)_{r=t_s}^{m+1}$ is a 0-mean martingale with respect to $(\mathcal{G}_r)_{r=t_s}^{m+1}$. First observe that $\tilde{A}_{t_s,r}$ is $\mathcal{G}_r$-measurable since the terms which compose it are $\mathcal{G}_r$-measurable, so M_r is $\mathcal{G}_r$-measurable.

Next, we show that $\mathbb{E}[M_{r+1} - M_r \mid \mathcal{G}_r] = 0$. When $r = m$ this follows by construction, so let $r < m$. Observe, importantly, that the scoring function and cap used to define $\mathcal{R}_{r+1}$ is $\mathcal{F}_r$-measurable, and since we get to see all of $(X_i)_{i=1}^n$ in $\mathcal{G}_r$ we in fact have that $\mathcal{R}_{r+1}$ is $\mathcal{G}_r$-measurable and also we can calculate how many points n_{r+1} fall within it. Then

$$M_{r+1} - M_r = \frac{1}{\sqrt{n_{t_s}}} \sum_{i \in \mathcal{I}_1 : X_i \in \mathcal{R}_r \backslash \mathcal{R}_{r+1}} (Y_{i,w} - \hat{Y}_{i,w})$$

$$= \frac{1}{\sqrt{n_{t_s}}} \sum_{i \in \mathcal{I}_1 : X_i \in \mathcal{R}_r \backslash \mathcal{R}_{r+1}} (g_1(X_i) - \hat{g}_1^{(1)}(X_i)) \left(1 - \frac{W_i}{e(X_i)}\right).$$

Note that n_{t_s}, $\mathcal{R}_r \setminus \mathcal{R}_{r+1}$, $(X_i)_{i=1}^n$, and $\hat{g}_1^{(1)}(\cdot)$ are $\mathcal{G}_r$-measurable. Then the only random quantities, conditional on $\mathcal{G}_r$, in the above expression, are the W_i. But as in the proof of Lemma E.4, conditioning on $\mathcal{G}_r$ does not affect the conditional distribution of $W_i \mid X_i$, and thus $1 - W_i/e(X_i)$ have mean 0 conditional on $\mathcal{G}_r$ and $\mathbb{E}[M_{r+1} - M_r \mid \mathcal{G}_r] = 0$. Now t_j is a bounded stopping time conditionally on $\mathcal{G}_{t_s}$, so we may apply the optional stopping theorem to conclude that $\mathbb{E}[M_{t_j} \mid \mathcal{G}_{t_s}] = 0$.

Now we consider the variance, which we decompose in terms of martingale differences as

$$\mathrm{Var}(M_{t_j} \mid \mathcal{G}_{t_s}) = \mathbb{E}\left[\sum_{r=t_s+1}^{t_j} \mathrm{Var}(M_{r+1} - M_r \mid \mathcal{G}_r) \,\Bigg|\, \mathcal{G}_{t_s}\right].$$

Then again recalling that conditionally on $\mathcal{G}_r$ the only random terms in the difference $M_{r+1} - M_r$ are the treatment indicators W_i, which are also independent,

$$\mathrm{Var}(M_{r+1} - M_r \mid \mathcal{G}_r) = \frac{1}{n_{t_s}} \sum_{i \in \mathcal{I}_1 : X_i \in \mathcal{R}_r \setminus \mathcal{R}_{r+1}} (g_1(X_i) - \hat{g}_1^{(1)}(X_i))^2 \mathrm{Var}\left(1 - \frac{W_i}{e(X_i)} \,\Big|\, \mathcal{G}_r\right)$$

$$\leq \frac{1}{n_{t_s}} \cdot L_{\mathrm{stable}}^2 \sum_{i \in \mathcal{I}_1 : X_i \in \mathcal{R}_r \setminus \mathcal{R}_{r+1}} \mathbb{E}\left[\left(1 - \frac{W_i}{e(X_i)}\right)^2 \,\Big|\, \mathcal{G}_r\right]$$

$$= \frac{1}{n_{t_s}} \cdot L_{\mathrm{stable}}^2 \sum_{i \in \mathcal{I}_1 : X_i \in \mathcal{R}_r \setminus \mathcal{R}_{r+1}} \frac{1 - e(X_i)}{e(X_i)}$$

$$\leq \frac{|\{i \in \mathcal{I}_1 : X_i \in \mathcal{R}_r \setminus \mathcal{R}_{r+1}\}|}{n_{t_s}} \cdot L_{\mathrm{stable}}^2 \cdot \frac{1 - \eta}{\eta}.$$

Therefore,

$$\sum_{r=t_s+1}^{t_j} \frac{|\{i \in \mathcal{I}_1 : X_i \in \mathcal{R}_r \setminus \mathcal{R}_{r+1}\}|}{n_{t_s}} \cdot L_{\mathrm{stable}}^2 \cdot \frac{1 - \eta}{\eta}$$

$$= \frac{L_{\mathrm{stable}}^2}{n_{t_s}} \frac{1 - \eta}{\eta} \sum_{r=t_s+1}^{t_j} |\{i \in \mathcal{I}_1 : X_i \in \mathcal{R}_r \setminus \mathcal{R}_{r+1}\}|$$

$$\leq \frac{L_{\mathrm{stable}}^2}{n_{t_s}} \frac{1 - \eta}{\eta} \cdot n_{t_s}$$

$$= L_{\mathrm{stable}}^2 \cdot \frac{1 - \eta}{\eta}.$$

We have bounded the sum in the second line by n_{t_s} by noticing that the sum is exactly adding up the number of points in-between $\mathcal{R}_{t_s+1}$ and $\mathcal{R}_{t_s+2}$, between $\mathcal{R}_{t_s+2}$ and $\mathcal{R}_{t_s+3}$, and so on. This can be no greater than n_{t_s}, the number of points in $\mathcal{R}_{t_s}$, which contains all of these regions. Then as $L_{\mathrm{stable}}^2(1 - \eta)/\eta$ is $\mathcal{G}_{t_s}$-measurable, we have

$$\mathrm{Var}(M_{t_j} \mid \mathcal{G}_{t_s}) \leq L_{\mathrm{stable}}^2 \cdot \frac{1 - \eta}{\eta} \xrightarrow{p} 0$$

by Assumption D.19, as desired. $\qquad \square$

With these facts, we are prepared to prove Lemma D.23.

Proof of Lemma D.23. Note that as we have established Lemma E.6, we may exactly imitate the proof of Lemma E.3 to obtain that

$$\sup_{t \in [m]_0 : \alpha_t > 0} |\hat{C}_t - C_t| \xrightarrow{p} 0.$$

Now we may write

$$\sup_{t \in [m]_0 : \alpha_t > 0} |d_t - \hat{d}_t| = \sup_{t \in [m]_0 : \alpha_t > 0} \left| C_t - \sqrt{\frac{\hat{V}_t}{V_t}} \cdot C_t - R_t \right|$$

$$\leq \sup_{t \in [m]_0 : \alpha_t > 0} \left| C_t - \sqrt{\frac{\hat{V}_t}{V_t}} \cdot C_t \right| + \sup_{t \in [m]_0 : \alpha_t > 0} |R_t|.$$

That the first supremum term in the final expression converges to 0 follows the proof of Lemma D.18 identically, where we note that this exact term appears in that proof. That the second supremum term converges to 0 follows from Lemma E.4. □

E.4 Distributional properties of chiseled regions

We work out some properties regarding the masses and conditional variances of chiseled regions. These are helpful for connecting some of the primitives described in Appendix D.4.7 to the general asymptotic conditions described in Appendix D.4.2. Some similar results can be found in Fraser (1951), though our proofs are self-contained.

In the following discussion we will allow the regions and filtrations to be defined for all positive integers t as in Equation (6). That is, we are working with $(\mathcal{R}_t)_{t=0}^\infty$ and $(\mathcal{F}_t)_{t=0}^\infty$. Recall, also, that $n_t := |\{i : X_i \in \mathcal{R}_t\}|$. For simplicity, we assume that Algorithm 2 chooses $c < \infty$ at most a finite number N of times, where N can be arbitrary, but is fixed. This way, $\mathcal{R}_t = \emptyset$ for all $t > n + N$.

Lemma E.7 (Joint distribution of region masses). *Suppose that $n_t - n_{t+1} = 1$ for all $t = 0, ..., n - 1$ almost surely. Then*

$$(\rho(\mathcal{R}_1), ..., \rho(\mathcal{R}_n)) \stackrel{d}{=} (U_{(n)}, ..., U_{(1)})$$

where $(U_{(1)}, ..., U_{(n)})$ are the joint order statistics of n independent standard uniforms.

Proof of Lemma E.7. Let $U_{m,n}$ be a random variable representing the mth order statistic of n uniform random variables. We will inductively characterize the distribution of $\rho(\mathcal{R}_t)$ in increasing order of t. In particular, we will show that for $t \geq 0$, $\rho(\mathcal{R}_{t+1}) \stackrel{d}{=} \rho(\mathcal{R}_t) \times U_{n-t,n-t}$ conditionally on $\mathcal{F}_t$. Since $\rho(\mathcal{R}_0) = 1$, we can thus sample $(\rho(\mathcal{R}_1), ..., \rho(\mathcal{R}_n))$ by independently generating $U_{n,n}, ..., U_{1,1}$ and setting $\rho(\mathcal{R}_t) = U_{n,n} \times ... \times U_{n-t,n-t}$ for each t. Lemma F.10 then directly shows that this is equivalent to sampling the order statistics of n independent uniforms, thus yielding the conclusion.

Let t be fixed, and let $f(\cdot)$ and c be what Algorithm 2 chooses based on $\mathcal{F}_t$. First define $\gamma := c \wedge \min_{i:X_i \in \mathcal{R}_t} f(X_i)$. Note that since we assume exactly one point is revealed with every application of chiseling, we must have that $\gamma = \min_{i:X_i \in \mathcal{R}_t} f(X_i)$ almost surely, and also that $f(X_i)$ has no ties almost surely conditionally on $\mathcal{F}_t$. Letting $F(\cdot)$ denote the (random, $\mathcal{F}_t$-measurable) CDF of $f(X) \mid X \in \mathcal{R}_t$, the latter condition implies that $F(Z)$ is uniformly distributed when Z is sampled from $X \mid X \in \mathcal{R}_t$. Now we implicitly condition on $\mathcal{F}_t$ and note that

$$
\begin{aligned}
\frac{\rho(\mathcal{R}_{t+1})}{\rho(\mathcal{R}_t)} &= \frac{\mathbb{P}(X \in \mathcal{R}_{t+1})}{\mathbb{P}(X \in \mathcal{R}_t)} \\
&= \mathbb{P}(X \in \mathcal{R}_{t+1} \mid X \in \mathcal{R}_t) \\
&= \mathbb{P}(Z \in \mathcal{R}_{t+1}) \\
&= \mathbb{P}(f(Z) > \gamma).
\end{aligned}
$$

To be clear: $\mathcal{R}_{t+1}$ and γ are still random conditionally on $\mathcal{F}_t$, and hence we read $\mathbb{P}(X \in \mathcal{R}_{t+1})$ as the random variable that arises from plugging the random variable $\mathcal{R}_{t+1}$ into the fixed function that evaluates $\mathbb{P}(X \in \mathcal{R})$, and so on. The first equality above is thus by definition and the second equality uses the fact that $\mathcal{R}_{t+1} \subseteq \mathcal{R}_t$ almost surely. The third equality is simply renaming; the fourth equality follows because $Z \in \mathcal{R}_t$ almost surely and hence belongs to $\mathcal{R}_{t+1}$ if and only if it additionally satisfies $f(Z) > \gamma$.

We have thus observed that the relative mass $\rho(\mathcal{R}_{t+1})/\rho(\mathcal{R}_t)$ is precisely the proportion of $f(Z)$'s mass that falls above γ; that is, $1 - F(\gamma)$. But because $F(\cdot)$ is non-decreasing,

$$
1 - F(\gamma) = 1 - \min_{i:X_i \in \mathcal{R}_t} F(f(X_i)) \overset{d}{=} 1 - \min_{j=1,\ldots,n-t} U_j \overset{d}{=} \max_{j=1,\ldots,n-t} U_j
$$

where U_j are i.i.d. uniform random variables. Above, we have applied Corollary 3.3 to conclude that $(f(X_i))_{i:X_i \in \mathcal{R}_t}$ are i.i.d. from the distribution $f(X) \mid X \in \mathcal{R}_t$ and hence $(F(f(X_i)))_{i:X_i \in \mathcal{R}_t}$ are i.i.d. uniforms. This shows us that conditionally on $\mathcal{F}_t$, $\rho(\mathcal{R}_{t+1})/\rho(\mathcal{R}_t)$ is distributed as $U_{n-t,n-t}$. Thus we have that $\rho(\mathcal{R}_{t+1}) = \rho(\mathcal{R}_t) \times (\rho(\mathcal{R}_{t+1})/\rho(\mathcal{R}_t)) \overset{d}{=} \rho(\mathcal{R}_t) \times U_{n-t,n-t}$ conditionally on $\mathcal{F}_t$ as desired. $\qquad\square$

Remark E.8 (Generalized CDF). *The proof of Lemma E.7 works in general for scores taking values in $\mathbb{R}^2$ with lexicographic ordering (Appendix C.3). We mention this because we will leverage a construction based on lexicographic order in the subsequent proofs which reference this lemma. Note the following generalization of a CDF F: if A is a random variable in $\mathbb{R}^2$ and $a \in \mathbb{R}^2$, define $F(a) := \mathbb{P}(A \leq_{\mathrm{lo}} a)$ where $\leq_{\mathrm{lo}}$ is lexicographic ordering. F retains the useful properties that (1) if $a \leq_{\mathrm{lo}} b$, then $F(a) \leq F(b)$, and (2) $F(A)$ has the standard uniform distribution as long as A is non-atomic, i.e. $\mathbb{P}(A = a) = 0$ for all a. Then $\leq_{\mathrm{lo}}$ may be used in place of the usual $\leq$ in the proof of Lemma E.7 without modification.*

Lemma E.7 characterizes the masses of the chiseled regions when they reveal one point at a time. We will soon prove a lemma that characterizes the distribution of masses more generally. But first we need a construction that allows us to "consolidate" a sequence of chiseled regions so that there is at least one point in-between every contiguous pair of non-empty regions, and one that allows us to "split" a sequence so that there is at most one point in-between every contiguous pair of regions. It is helpful to first offer an intuitive

picture. One way to view chiseling is as a continuous process of shrinking the covariate space which reacts to data points as they are excluded from the region. The sequence of regions $(\mathcal{R}_t)_{t=0}^{\infty}$ are simply discrete "snapshots" of this continuous process, where the choice of whether to take a snapshot at any particular point depends only on the data points that have been excluded from the current region. Intuitively, then, we may construct instantiations of chiseling that correspond to fundamentally the same shrinking process but which take different snapshots—perhaps recording more intermediate stages of the shrinking process, or skipping over some previously recorded regions. The following lemma simply states that we may choose to not snapshot regions that have not revealed any new data points since the last snapshotted region. The lemma which follows states that we may snapshot more often so that each snapshot captures a change in at most one data point of information (in an augmented covariate space).

In the below, by an "instantiation of chiseling," we mean a way of choosing $f(\cdot)$ and c in Algorithm 2 from the data. This way of choosing is itself a function from data, outputting the pair $(f(\cdot), c)$.

Lemma E.9 (Consolidating regions)**.** *Let $(\mathcal{R}_t)_{t=0}^{\infty}$ be the sequence of regions produced by a particular instantiation of chiseling. Also, for some fixed $N < \infty$, suppose that in this instantiation, Algorithm 2 sets $c < \infty$ at most N times almost surely. Then there exists an instantiation of chiseling that produces a subsequence of regions $(\tilde{\mathcal{R}}_t)_{t=0}^{\infty} \subseteq (\mathcal{R}_t)_{t=0}^{\infty}$ with the property that for all $t \geq 0$, either $\{i : X_i \in \tilde{\mathcal{R}}_{t-1} \setminus \tilde{\mathcal{R}}_t\}$ is nonempty or $\tilde{\mathcal{R}}_t = \emptyset$ almost surely. Furthermore, we have that almost surely, for all t such that $\{i : X_i \in \mathcal{R}_{t-1} \setminus \mathcal{R}_t\}$ is nonempty, $\mathcal{R}_t \in (\tilde{\mathcal{R}}_t)_{t=0}^{\infty}$.*

Proof of Lemma E.9. Let us index the scoring functions as $f_t(\cdot)$ and caps as c_t; that is, the first application of chiseling will use $f_1(\cdot)$, and c_1 in Algorithm 2, and so on. Suppose that r is the smallest index such that $c_r < \infty$; note that r is random, a stopping time, and that we may have $r = \infty$. We describe a modified instantiation of chiseling that uses the scoring functions $\tilde{f}_t(\cdot)$ and caps $\tilde{c}_t$, producing the regions $(\tilde{\mathcal{R}}_t)_{t=0}^{\infty}$. For all $t < r$, we let $\tilde{f}_t(\cdot) = f_t(\cdot)$ and $\tilde{c}_t = c_t$. If $r = \infty$, then this corresponds to no modification. Otherwise if $r < \infty$, let $f'(\cdot)$ and c' be the values that $f_{r+1}(\cdot)$ and c_{r+1} *would* take in the event that $\{i : X_i \in \mathcal{R}_{r-1} \setminus \mathcal{R}_r\}$ is empty. Note that both $f'(\cdot)$ and c' can be determined without actually having to observe whether $\{i : X_i \in \mathcal{R}_{r-1} \setminus \mathcal{R}_r\}$ is empty, i.e. they are $\mathcal{F}_r$-measurable (i.e. the analyst can "think ahead"). Then we let

$$\tilde{f}_r(x) = \begin{cases} \arctan(f_r(x)) & \text{if } f_r(x) \leq c_r, \\ \pi + \arctan(f'(x)) & \text{if } f_r(x) > c_r, \end{cases} \quad \text{and} \quad \tilde{c}_r = \begin{cases} \infty & \text{if } c' = \infty, \\ \pi + \arctan(c') & \text{otherwise.} \end{cases}$$

Straightforward, though tedious, algebra and case-checking verifies that $\tilde{\mathcal{R}}_r = \mathcal{R}_r$ if $\{i : X_i \in \mathcal{R}_{r-1} \setminus \mathcal{R}_r\}$ is nonempty and $\tilde{\mathcal{R}}_r = \mathcal{R}_{r+1}$ otherwise. Intuitively, all the above is doing is shrinking according to $f_r(\cdot)$ as long as we are below the cap c_r, and switching to $f'(\cdot)$ thereafter, stopping at c'. The role of $\arctan(\cdot)$ and the shifting by π is simply to separate the two scores into non-overlapping ranges so that we can shrink according to one scoring function followed by the other.

Finally, for all $t > r$ we define $\tilde{f}_t(\cdot) = f_t(\cdot)$ and $\tilde{c}_t = c_t$ if $\{i : X_i \in \mathcal{R}_{r-1} \setminus \mathcal{R}_r\}$ is nonempty, and $\tilde{f}_{t+1}(\cdot) = f_{t+1}(\cdot)$ and $\tilde{c}_{t+1} = c_{t+1}$ otherwise. Note that this last construction

only makes sense if $\tilde{f}_t(\cdot)$ and $\tilde{c}_t(\cdot)$ can be calculated from the revealed data points in the modified instantiation. But this is indeed the case, since by the above we have access to all the data points outside of $\mathcal{R}_r$ when $\{i : X_i \in \mathcal{R}_{r-1} \setminus \mathcal{R}_r\}$ is nonempty, and all the data points outside of $\mathcal{R}_{r+1}$ otherwise. Thus, the information necessary to determine $f_{r+1}(\cdot)$ in the former case and $f_{r+2}(\cdot)$ in the latter case is available. It follows that $\tilde{\mathcal{R}}_{r+1} = \mathcal{R}_{r+1}$ in the former case and $\tilde{\mathcal{R}}_{r+1} = \mathcal{R}_{r+2}$ in the latter case. Then applying this argument inductively, for all $t > r$ we have that $\tilde{f}_t(\cdot)$ and $\tilde{c}_t$ can be calculated from the revealed information (i.e. this is a valid instantiation of chiseling), and that $\tilde{\mathcal{R}}_t = \mathcal{R}_t$ when $\{i : X_i \in \mathcal{R}_{r-1} \setminus \mathcal{R}_r\}$ is nonempty and $\tilde{\mathcal{R}}_t = \mathcal{R}_{t+1}$ otherwise.

Thus, we have described a valid instantiation of chiseling that produces regions $(\tilde{\mathcal{R}}_t)_{t=0}^{\infty} \subseteq (\mathcal{R}_t)_{t=0}^{\infty}$. Furthermore, the modified instantiation sets $\tilde{c} < \infty$ in Algorithm 2 at most $N-1$ times almost surely, because the modified instantiation almost surely sets one fewer finite cap than the original instantiation (in particular, cutting out the finite cap at stage r) as long as the latter sets at least one finite cap. Lastly, it preserves all regions $\mathcal{R}_t$ where $\{i : X_i \in \mathcal{R}_{t-1} \setminus \mathcal{R}_t\}$ is nonempty by construction, since the only possibility is that it removes $\mathcal{R}_r$ if $\mathcal{R}_r$ does not satisfy this property.

By again modifying the modified instantiation using the same construction as above, we can obtain an instantiation that sets at most $N-2$ finite caps, and so on until we obtain an instantiation that sets zero finite caps. Each modified instantiation produces regions that are subsequences of the regions outputted by the instantiation it modifies, and never does it remove a region $\mathcal{R}_t$ with the property that $\{i : X_i \in \mathcal{R}_{t-1} \setminus \mathcal{R}_t\}$ is nonempty. Since in the final modification all caps are infinite, we must have that each application of chiseling reveals at least one point as long as there are points left. That is, if $(\tilde{\mathcal{R}}_t)_{t=0}^{\infty} \subseteq (\mathcal{R}_t)_{t=0}^{\infty}$ denotes the regions produced by the final modified instantiation, then $\{i : X_i \in \tilde{\mathcal{R}}_{t-1} \setminus \tilde{\mathcal{R}}_t\}$ is nonempty as long as $\mathcal{R}_t \neq \emptyset$, as desired. $\qquad\square$

Lemma E.10 (Splitting regions). *Let $(\mathcal{R}_t)_{t=0}^{\infty}$ be the sequence of regions produced by a particular instantiation of chiseling. Also, for some fixed $N < \infty$, suppose that in this instantiation, Algorithm 2 sets $c < \infty$ at most N times almost surely. Let $U_1, ..., U_n$ be i.i.d. standard uniforms, and define $\tilde{X}_i := (X_i, U_i)$ for $i = 1, ..., n$. Then there exists an instantiation of chiseling applied to $(\tilde{X}_i, Y_i)_{i=1}^n$ that produces a supersequence of augmented regions $(\tilde{\mathcal{R}}_t)_{t=0}^{\infty} \supseteq (\mathcal{R}_t \times [0,1])_{t=0}^{\infty}$ with the property that $\{i : \tilde{X}_i \in \tilde{\mathcal{R}}_{t-1} \setminus \tilde{\mathcal{R}}_t\}$ has at most one element almost surely for all $t \geq 0$.*

Proof of Lemma E.10. In this proof we use the tiebreaking construction described in Appendix C.5. Thus, orderings here should be understood lexicographically. We constructively describe an instantiation of chiseling and the chiseled regions $\tilde{\mathcal{R}}_t$. First, we note that $\tilde{\mathcal{R}}_0 = \mathcal{X} \times [0,1] = \mathcal{R}_0 \times [0,1]$ by construction. We describe a sequence of chiseled regions $\tilde{\mathcal{R}}_1, ..., \tilde{\mathcal{R}}_k$ of variable length k such that $\tilde{\mathcal{R}}_k$ coincides with $\mathcal{R}_1 \times [0,1]$ and $\tilde{\mathcal{R}}_{k-1} \setminus \tilde{\mathcal{R}}_k$ contains at most one data point. As in the proof of Lemma E.9, let us index the scoring functions in the original instantiation as $f_t(\cdot)$ and caps as c_t. In the modified instantiation, run Algorithm 2 with

$$\tilde{f}(x, u) = (f_1(x), u) \quad \text{and} \quad \tilde{c} = (c_1, 1)$$

once, and let the outputted region be $\tilde{\mathcal{R}}_1$. If $\{i : \tilde{X}_i \in \tilde{\mathcal{R}}_0 \setminus \tilde{\mathcal{R}}_1\}$ is empty, then $\tilde{\mathcal{R}}_1 = \mathcal{R}_1 \times [0,1]$. Otherwise, let $\gamma = \min_{i:(X_i, U_i) \in \tilde{\mathcal{R}}_0 \setminus \tilde{\mathcal{R}}_1} f_1(X_i)$ and repeatedly run Algorithm 2 with the same

$\tilde{f}(x, u)$ as above but with $\tilde{c} = (\gamma, 1)$ until no new points are revealed, which must happen eventually since there are finitely many data points. That is, let $\tilde{\mathcal{R}}_2, ..., \tilde{\mathcal{R}}_k$ be obtained by chiseling with the above scoring function and cap and such that $\{i : \tilde{X}_i \in \tilde{\mathcal{R}}_{t-1} \setminus \tilde{\mathcal{R}}_t\}$ has exactly one element for $t = 3, ..., k-1$ (since by using the auxiliary dimension we encounter no ties almost surely) and $\{i : \tilde{X}_i \in \tilde{\mathcal{R}}_{k-1} \setminus \tilde{\mathcal{R}}_k\}$ is empty. Then

$$\begin{aligned}
\tilde{\mathcal{R}}_k &= \tilde{\mathcal{R}}_0 \cap \{(x, u) \in \mathcal{X} \times [0, 1] : (f_1(x), u) > (\gamma, 1)\} \\
&= \mathcal{R}_0 \cap \{(x, u) \in \mathcal{X} \times [0, 1] : f_1(x) > \gamma\} \\
&= \mathcal{R}_1 \times [0, 1].
\end{aligned}$$

The first equality follows since $\tilde{\mathcal{R}}_1, ..., \tilde{\mathcal{R}}_k$ are chiseled using the same $\tilde{f}(x, u)$ and thus it suffices to consider just the most stringent threshold for $\tilde{f}(x, u)$, which must be $(\gamma, 1)$ since that is the cap and we chisel until no new points are revealed. Now since we have revealed the same data points that would have been revealed in the original instantiation after shrinking to $\mathcal{R}_1$, we can deduce $f_2(\cdot)$ and c_2 from this information in the modified instantiation. Then we may repeat this construction to obtain a further sequence of regions $\tilde{\mathcal{R}}_{k+1}, ..., \tilde{\mathcal{R}}_{k+l}$ such that $\tilde{\mathcal{R}}_{k+l} = \mathcal{R}_2 \times [0, 1]$ almost surely. Recursively repeating this construction results in a supersequence $(\tilde{\mathcal{R}}_t)_{t=0}^{\infty}$ of $(\mathcal{R}_t \times [0, 1])_{t=0}^{\infty}$. Also, we have already seen that since we use the auxiliary dimension to tiebreak that we must reveal at most one element with every application of Algorithm 2, i.e. $\{i : \tilde{X}_i \in \tilde{\mathcal{R}}_{t-1} \setminus \tilde{\mathcal{R}}_t\}$ has at most one element as desired. $\square$

Now we state a general characterization of the masses of chiseled regions.

Lemma E.11 (Distribution of smallest mass with sufficient samples). *Let $1 < m \leq n$ be fixed. Define*

$$L_{\text{vol}}^{-1} := \inf_{t \geq 0 : n_t \geq m} \rho(\mathcal{R}_t).$$

Let $U_{m,n}$ be a random variable representing the mth order statistic of n uniform random variables. Then L_{vol}^{-1} stochastically dominates $U_{m,n}$.

Proof of Lemma E.11. We will construct a dataset $(\tilde{X}_i, Y_i)_{i=1}^n$ coupled to the same probability space as $(X_i, Y_i)_{i=1}^n$ and a sequence of chiseled regions $(\tilde{\mathcal{R}}_t)_{t=0}^{\infty}$ obtained from the coupled dataset with certain properties. First, define $\tilde{n}_t := |\{i : \tilde{X}_i \in \tilde{\mathcal{R}}_t\}|$ and $\tilde{\rho}(\tilde{\mathcal{R}}) := \mathbb{P}(\tilde{X} \in \tilde{\mathcal{R}})$. We will show that

$$L_{\text{vol}}^{-1} = \inf_{t \geq 0 : n_t \geq m} \rho(\mathcal{R}_t) \geq \inf_{t \geq 0 : \tilde{n}_t \geq m-1} \tilde{\rho}(\tilde{\mathcal{R}}_t) \quad \text{a.s.} \tag{28}$$

and that $\tilde{n}_t - \tilde{n}_{t+1} = 1$ for all $t = 0, ..., n-1$ almost surely. Then as $\tilde{n}_t = n - t$, the minimum is attained when $m - 1 = n - t$, or equivalently when $t = n - m + 1$. Thus, the right-hand side further reduces to $\tilde{\rho}(\tilde{\mathcal{R}}_{n-m+1})$, and since this term arises from an instantiation of chiseling, by Lemma E.7 it is distributed as $U_{m,n}$, and this will complete the proof.

Let $(\tilde{X}_i, Y_i)_{i=1}^n$ be as in the statement of Lemma E.10, and let $(\mathcal{R}'_t)_{t=0}^{\infty}$ be the supersequence of regions given by the lemma. Then let $(\tilde{\mathcal{R}}_t)_{t=0}^{\infty}$ be the subsequence of $(\mathcal{R}'_t)_{t=0}^{\infty}$ obtained by consolidating the latter using Lemma E.9. Since $(\mathcal{R}'_t)_{t=0}^{\infty}$ is such that each successive region reveals at most one data point, and $(\tilde{\mathcal{R}}_t)_{t=0}^{\infty}$ is precisely the subsequence that skips over regions that do not reveal any new data points, we have that $\{i : \tilde{X}_i \in \tilde{\mathcal{R}}_{t-1} \setminus \tilde{\mathcal{R}}_t\}$ has exactly one element for all $t = 1, ..., n$. That is, $\tilde{n}_{t-1} - \tilde{n}_t = 1$ for all $t = 1, ..., n$. Furthermore, we note that for every t, there is a $\tilde{\mathcal{R}}_{s(t)}$ such that $\mathcal{R}_t \times [0, 1] \supseteq \tilde{\mathcal{R}}_{s(t)}$ and $n_t - \tilde{n}_{s(t)} \leq 1$. That is,

for every region $\mathcal{R}_t$ there is a region $\tilde{\mathcal{R}}_{s(t)}$ contained in $\mathcal{R}_t \times [0,1]$ that excludes at most one additional point. To see why, note that $\mathcal{R}_t \times [0,1] = \mathcal{R}'_u$ for some u. Let s be the smallest integer such that $\tilde{\mathcal{R}}_s \subseteq \mathcal{R}'_u$. Then $n_t - \tilde{n}_s \leq 1$, since splitting creates a sequence that reveals at most one data point at each stage, consolidation only excludes regions that revealed no data points since the last stage, and $(\tilde{\mathcal{R}}_t)_{t=0}^\infty$ is obtained by consolidating $(\mathcal{R}'_t)_{t=0}^\infty$, which is obtained by splitting.

Finally, we must show the domination in Equation (28). But this simply follows because if $n_t \geq m$, then $\tilde{n}_{s(t)} \geq m-1$ and $\rho(\mathcal{R}_t) = \tilde{\rho}(\mathcal{R}_t \times [0,1]) \geq \tilde{\rho}(\tilde{\mathcal{R}}_{s(t)})$, and hence for every element in the left-hand side infimum there is a smaller element in the right-hand side infimum. $\qquad\square$

This result allows us to easily establish the following corollary.

Corollary E.12 (L^1-bound on inverse volume). *Let $p_{\mathrm{prop}} \in (0,1)$ be fixed, and let*

$$L_{\mathrm{vol}} := \sup_{t \geq 0 : n_t \geq p_{\mathrm{prop}} \cdot n} \rho(\mathcal{R}_t)^{-1}.$$

Then as long as $p_{\mathrm{prop}} \cdot n > 1$,

$$\mathbb{E}[L_{\mathrm{vol}}] \leq \frac{1}{p_{\mathrm{prop}} - n^{-1}}.$$

In particular, $\mathbb{E}[L_{\mathrm{vol}}] \leq 2/p_{\mathrm{prop}}$ for $n > 2/p_{\mathrm{prop}}$.

Proof of Corollary E.12. Letting $m = \lceil p_{\mathrm{prop}} \cdot n \rceil$, Lemma E.11 tells us that L_{vol} is stochastically dominated by $U_{m,n}^{-1}$. We note that $U_{m,n} \sim \mathrm{Beta}(m, n-m+1)$, which has the same distribution as $Z_1/(Z_1 + Z_2)$ where $Z_1 \sim \mathrm{Gamma}(m,1)$ and $Z_2 \sim \mathrm{Gamma}(n-m+1,1)$ are independent. Thus, $U_{m,n}^{-1}$ has the same distribution as $(Z_1 + Z_2)/Z_1 = 1 + Z_2/Z_1 = 1 + B'$ where $B' \sim \mathrm{BetaPrime}(n-m+1, m)$. The mean of the beta prime distribution is a standard fact and in this case given by $\frac{n-m+1}{m-1}$ (note that $m > 1$), so

$$\mathbb{E}[L_{\mathrm{vol}}] \leq 1 + \frac{n-m+1}{m-1} = \frac{n}{m-1} \leq \frac{n}{p_{\mathrm{prop}} \cdot n - 1} = \frac{1}{p_{\mathrm{prop}} - n^{-1}}.$$

The right-hand side is bounded by $2/p_{\mathrm{prop}}$ for $n > 2/p_{\mathrm{prop}}$. $\qquad\square$

The bound on the inverse volume helps us establish another useful bound on the smallest conditional variance of any tested region in the pointwise setting. The following lemma reintroduces dependence on n, since we will show that a certain random sequence is bounded in probability. As in Appendix D.4, formal statements will include n superscripts where appropriate but otherwise we will suppress dependence on n for readability.

Lemma E.13 (Probability bound on inverse of smallest variance). *Suppose that we work under the pointwise asymptotics defined in Section 3.6. Let $p_{\mathrm{prop}} \in (0,1)$ be fixed and define*

$$L_{\mathrm{vmin}}^{(n)} := \left(\inf_{t \geq 0 : n_t^{(n)} \geq p_{\mathrm{prop}} \cdot n} \sigma_1(\mathcal{R}_t^{(n)}) \right)^{-1}.$$

Then under Assumption 3.11, $L_{\mathrm{vmin}}^{(n)}$ is bounded in probability.

Proof of Lemma E.13. Let us define

$$B(q) := \inf_{\mathcal{R} \subseteq \mathcal{X}: \rho(\mathcal{R}) \geq q} \sigma_1(\mathcal{R}).$$

Note that Assumption 3.11 implies that $B(q)^2 > 0$ for all $0 < q \leq 1$ (Lemma F.11). Hence, $B(q) > 0$ for all $0 < q \leq 1$.

Let $\epsilon > 0$. Recall the definition of L_{vol} from Corollary E.12 and note that

$$L_{\text{vol}} \leq q^{-1} \implies \inf_{t \geq 0: n_t \geq p_{\text{prop}} \cdot n} \rho(\mathcal{R}_t) \geq q$$
$$\implies \inf_{t \geq 0: n_t \geq p_{\text{prop}} \cdot n} \sigma_1(\mathcal{R}_t) \geq B(q)$$
$$\implies L_{\text{vmin}} \leq B(q)^{-1}.$$

Since L_{vol} is L^1-bounded by Corollary E.12 (where we have now made L_{vol} dependent upon n), it is also bounded in probability. So we may choose $q > 0$ such that $\mathbb{P}(L_{\text{vol}} > q^{-1}) < \epsilon$ for sufficiently large n. Then

$$\mathbb{P}(L_{\text{vmin}} > B(q)^{-1}) \leq \mathbb{P}(L_{\text{vol}} > q^{-1}) < \epsilon$$

for sufficiently large n, and hence L_{vmin} is bounded in probability. $\qquad\square$

Appendix F   Proofs of technical lemmas

The proofs in Appendix F are self-contained. It will help to consider the notational choices used here as independent of those made in the rest of the paper.

F.1   Measure theory refreshers

Lemma F.1 (Nonzero conditional probability if event happens)**.** *Let A be an event and let $\mathcal{F}$ be a σ-algebra. Then*

$$\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}] > 0 \ \textit{a.s. conditional on } A.$$

Proof of Lemma F.1. Let $Z := \mathbb{1}\{\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}] = 0\}$. Note that Z is $\mathcal{F}$-measurable. Then

$$0 = \mathbb{E}[0] = \mathbb{E}[\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}] \cdot Z] = \mathbb{E}[\mathbb{E}[\mathbb{1}\{A\} \cdot Z \mid \mathcal{F}]] = \mathbb{E}[\mathbb{1}\{A\} \cdot Z].$$

Then $\mathbb{1}\{A\} \cdot Z = 0$ almost surely, since it is nonnegative with expectation 0. Then for almost all $\omega \in A$, this implies that

$$0 = \mathbb{1}\{A\}(\omega) \cdot Z(\omega) = Z(\omega) \implies \mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}](\omega) > 0$$

since $\mathbb{1}\{A\}(\omega) = 1$. $\qquad\square$

Lemma F.2 (Measure-theoretic conditional expectation)**.** *Let X be a random variable with finite first moment, A be an event, and $\mathcal{F}$ be a σ-algebra. Then*

$$\mathbb{E}[X \mid \mathcal{F}, A] = \mathbb{E}[X \mid \mathcal{F}, \mathbb{1}\{A\}] \quad \textit{a.s. conditional on } A.$$

Proof of Lemma F.2. Recall that we have defined

$$\mathbb{E}[X \mid \mathcal{F}, A] = \frac{\mathbb{E}[X \mathbb{1}\{A\} \mid \mathcal{F}]}{\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}]}$$

where originally we have taken the convention that $0/0$ is undefined. Note that since $\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}] > 0$ almost surely conditional on A (Lemma F.1), then for the purposes of this proof, we may as well take the convention $0/0 = 0$, since these two conventions result in definitions that agree almost everywhere on A. Thus, we subsequently think of $\mathbb{E}[X \mid \mathcal{F}, A]$ as being well-defined almost everywhere. This will be a much more convenient object to work with. Now note that

$$\frac{\mathbb{E}[X \mathbb{1}\{A\} \mid \mathcal{F}]}{\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}]} \cdot \mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}] = \mathbb{E}[X \mathbb{1}\{A\} \mid \mathcal{F}]$$

since when $\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}] = 0$, the left-hand side evaluates to $0 \cdot 0 = 0$ and the right-hand side evaluates to 0. Otherwise, the denominator cancels ordinarily.

Next, note that $\mathcal{P} := \{S_1 \cap S_2 : S_1 \in \mathcal{F}, S_2 \in \{A, A^c, \Omega\}\}$ is a π-system and $\sigma(\mathcal{P}) = \sigma(\mathcal{F}, \mathbb{1}\{A\})$. We will check that

$$\mathbb{E}\left[\frac{\mathbb{E}[X \mathbb{1}\{A\} \mid \mathcal{F}]}{\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}]} \cdot \mathbb{1}\{A\} \cdot \mathbb{1}\{S\}\right] = \mathbb{E}[X \mathbb{1}\{A\} \cdot \mathbb{1}\{S\}] \tag{29}$$

for all $S \in \mathcal{P}$. But probability measures are uniquely defined by their values on π-systems, and thus the above will hold for all $S \in \sigma(\mathcal{F}, \mathbb{1}\{A\})$. Hence,

$$\mathbb{E}[X \mid \mathcal{F}, \mathbb{1}\{A\}] \cdot \mathbb{1}\{A\} = \mathbb{E}[X \mathbb{1}\{A\} \mid \mathcal{F}, \mathbb{1}\{A\}] = \frac{\mathbb{E}[X \mathbb{1}\{A\} \mid \mathcal{F}]}{\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}]} \cdot \mathbb{1}\{A\} \quad \text{a.s.}$$

by the measure-theoretic definition and uniqueness of conditional expectation. But this is equivalent to the statement in the lemma. Now we check Equation (29). Below, $S_1 \in \mathcal{F}$.

1. **Case 1:** $S = S_1 \cap A$. Then

$$\begin{aligned}
\mathbb{E}\left[\frac{\mathbb{E}[X \mathbb{1}\{A\} \mid \mathcal{F}]}{\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}]} \cdot \mathbb{1}\{A\} \cdot \mathbb{1}\{S\}\right] &= \mathbb{E}\left[\frac{\mathbb{E}[X \mathbb{1}\{A\} \mid \mathcal{F}]}{\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}]} \cdot \mathbb{1}\{A\} \cdot \mathbb{1}\{S_1\} \cdot \mathbb{1}\{A\}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{E}[X \mathbb{1}\{A\} \mid \mathcal{F}]}{\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}]} \cdot \mathbb{1}\{A\} \cdot \mathbb{1}\{S_1\} \Big| \mathcal{F}\right]\right] \\
&= \mathbb{E}\left[\frac{\mathbb{E}[X \mathbb{1}\{A\} \mid \mathcal{F}]}{\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}]} \cdot \mathbb{1}\{S_1\} \cdot \mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}]\right] \\
&= \mathbb{E}[\mathbb{E}[X \mathbb{1}\{A\} \mid \mathcal{F}] \cdot \mathbb{1}\{S_1\}] \\
&= \mathbb{E}[\mathbb{E}[X \mathbb{1}\{A\} \cdot \mathbb{1}\{S_1\} \mid \mathcal{F}]] \\
&= \mathbb{E}[X \mathbb{1}\{A\} \cdot \mathbb{1}\{S_1\}] \\
&= \mathbb{E}[X \mathbb{1}\{A\} \cdot \mathbb{1}\{S\}]
\end{aligned}$$

where to go to the last line we use the fact that $\mathbb{1}\{A\} \cdot \mathbb{1}\{S_1\} = \mathbb{1}\{A\} \cdot \mathbb{1}\{S\}$.

115

2. **Case 2:** $S = S_1 \cap A^c$. Then

$$\mathbb{E}\left[\frac{\mathbb{E}[X\mathbb{1}\{A\} \mid \mathcal{F}]}{\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}]} \cdot \mathbb{1}\{A\} \cdot \mathbb{1}\{S\}\right] = \mathbb{E}\left[\frac{\mathbb{E}[X\mathbb{1}\{A\} \mid \mathcal{F}]}{\mathbb{E}[\mathbb{1}\{A\} \mid \mathcal{F}]} \cdot \mathbb{1}\{A\} \cdot \mathbb{1}\{S_1\} \cdot \mathbb{1}\{A^c\}\right] = 0$$

and $\mathbb{E}[X\mathbb{1}\{A\} \cdot \mathbb{1}\{S\}] = \mathbb{E}[X\mathbb{1}\{A\} \cdot \mathbb{1}\{S_1\} \cdot \mathbb{1}\{A^c\}] = 0$.

3. **Case 3:** $S = S_1 \cap \Omega = S_1$. Then we may take the exact same calculation and reasoning as in **Case 1**.

$\square$

Lemma F.3 (Conditional equivalence at locally invertible points). *Let X and Z be random variables where Z is discrete. Let h be a measurable function and define $A := \{\omega \in \Omega : h^{-1}(h(Z(\omega)))$ is a singleton$\}$. Then*

$$\mathbb{E}[X \mid Z] = \mathbb{E}[X \mid h(Z)] \quad a.s. \text{ conditional on } A.$$

Proof of Lemma F.3. This follows directly from

$$\mathbb{E}[X \mid h(Z)](\omega) = \mathbb{E}[X \mid h(Z) = h(Z(\omega))] = \mathbb{E}[X \mid Z = Z(\omega)] = \mathbb{E}[X \mid Z](\omega)$$

because $h(Z) = h(Z(\omega)) \iff Z = Z(\omega)$ for $\omega \in \{\omega \in \Omega : h^{-1}(h(Z(\omega)))$ is a singleton$\}$. $\square$

Lemma F.4 (Stopped conditional expectation evaluated at a point). *Let Z be a bounded random variable and $(\mathcal{F}_t)_{t=0}^\infty$ be a filtration. Let ν be an almost surely finite stopping time with respect to $(\mathcal{F}_t)_{t=0}^\infty$ and define the stopped σ-algebra $\mathcal{F}_\nu := \{A : \{\nu \le t\} \cap A \in \mathcal{F}_t \text{ for all } t\}$. Then*

$$\mathbb{E}[Z \mid \mathcal{F}_\nu] = \sum_{t=0}^\infty \mathbb{E}[Z \mid \mathcal{F}_t]\mathbb{1}\{\nu = t\} \quad a.s.$$

Proof of Lemma F.4. Write the sum on the right-hand side as M. We first show that M is $\mathcal{F}_\nu$-measurable. Let $B \subseteq \mathbb{R}$ be Borel measurable. We will show that $\{M \in B\} \in \mathcal{F}_\nu$, or equivalently that

$$\{\nu \le t\} \cap \{M \in B\} \in \mathcal{F}_t \quad \text{for all } t.$$

But this follows since conditional on $\{\nu \le t\}$, we have $M = \sum_{s=0}^t \mathbb{E}[Z \mid \mathcal{F}_s]\mathbb{1}\{\nu = s\}$. Each term in the sum is an $\mathcal{F}_t$-measurable quantity, so conditional on $\{\nu \le t\}$, M is $\mathcal{F}_t$-measurable.

Now let $A \in \mathcal{F}_\nu$. Note that $\{\nu \le t\} \cap A$ and $\{\nu \le t-1\} \cap A$ are both $\mathcal{F}_t$-measurable by definition (the latter since $\mathcal{F}_{t-1} \subseteq \mathcal{F}_t$). Taking the difference implies that $\{\nu = t\} \cap A$ is $\mathcal{F}_t$-measurable. Then

$$\mathbb{E}[Z\mathbb{1}\{A\}] = \sum_{t=0}^\infty \mathbb{E}[Z\mathbb{1}\{A\}\mathbb{1}\{\nu = t\}]$$

$$= \sum_{t=0}^\infty \mathbb{E}[\mathbb{E}[Z\mathbb{1}\{A\}\mathbb{1}\{\nu = t\} \mid \mathcal{F}_t]]$$

$$= \sum_{t=0}^\infty \mathbb{E}[\mathbb{1}\{A\}\mathbb{1}\{\nu = t\}\mathbb{E}[Z \mid \mathcal{F}_t]]$$

$$= \mathbb{E}\left[\mathbb{1}\{A\}\sum_{t=0}^\infty \mathbb{1}\{\nu = t\}\mathbb{E}[Z \mid \mathcal{F}_t]\right].$$

116

The first equality follows since $\sum_{t=0}^{\infty} \mathbb{1}\{\nu = t\} = 1$ as ν is almost surely finite. The second equality follows from the tower rule, the third equality follows since $\mathbb{1}\{A\}\mathbb{1}\{\nu = t\} = \mathbb{1}\{A \cap \{\nu = t\}\}$ which is $\mathcal{F}_t$-measurable, so the product can be pulled out of the expectation. The last equality simply pushes the sum back inside the expectation. The result then follows by the uniqueness of conditional expectation. $\qquad\square$

F.2 Proof of Lemma D.1

Proof. Let Ω denote the sample space of the base probability measure. Let $\omega \in \Omega$ be fixed. Then

$$\gamma = \gamma(\omega) \text{ and } \mathcal{I} = \mathcal{I}(\omega) \iff S_i = \gamma(\omega) \text{ for all } i \notin \mathcal{I}(\omega) \text{ and } S_i > \gamma(\omega) \text{ for all } i \in \mathcal{I}(\omega).$$

Hence, for any measurable $B \in \mathbb{R}^{|\mathcal{I}(\omega)|}$,

$$
\begin{aligned}
&\mathbb{P}((S_i)_{i \in \mathcal{I}} \in B \mid \gamma, \mathcal{I})(\omega) \\
=&\mathbb{P}((S_i)_{i \in \mathcal{I}(\omega)} \in B \mid \gamma = \gamma(\omega), \mathcal{I} = \mathcal{I}(\omega)) \\
=&\mathbb{P}((S_i)_{i \in \mathcal{I}(\omega)} \in B \mid S_i = \gamma(\omega) \text{ for all } i \notin \mathcal{I}(\omega), S_i > \gamma(\omega) \text{ for all } i \in \mathcal{I}(\omega)) \\
=&\mathbb{P}((S_i)_{i \in \mathcal{I}(\omega)} \in B \mid S_i > \gamma(\omega) \text{ for all } i \in \mathcal{I}(\omega)).
\end{aligned}
$$

The last line follows because $(S_i)_{i \in \mathcal{I}(\omega)}$ is independent of $(S_i)_{i \notin \mathcal{I}(\omega)}$; note here that $\mathcal{I}(\omega)$ is a fixed set. Also, the last line exactly describes the measure of $|\mathcal{I}(\omega)|$ draws from the distribution of $S \mid S > \gamma(\omega)$. $\qquad\square$

F.3 Convergences, inequalities, and distributional facts

Lemma F.5 (Discrete conditional stochastic dominance). *Let X, Y be discrete random variables on a common support $[n] := \{0, ..., n\}$. Denote the PMF and CDF of X by f and F, and the PMF and CDF of Y by g and G, respectively. Then*

$$\frac{F(c)}{F(\gamma)} \leq \frac{G(c)}{G(\gamma)} \quad \text{for all } c \leq \gamma \in [n] \iff \frac{f(k)}{F(k)} \geq \frac{g(k)}{G(k)} \quad \text{for all } k \in [n].$$

Proof of Lemma F.5. Since $\frac{F(c)}{F(\gamma)} \leq \frac{G(c)}{G(\gamma)} \iff \frac{F(c)}{G(c)} \leq \frac{F(\gamma)}{G(\gamma)}$, the first statement is equivalent to the claim that $F(k)/G(k)$ is increasing in k. But this holds if and only if for all $k \in [n]$,

$$
\begin{aligned}
\frac{F(k)}{G(k)} - \frac{F(k-1)}{G(k-1)} \geq 0 &\iff \frac{F(k)}{G(k)} \geq \frac{F(k-1)}{G(k-1)} \\
&\iff \frac{G(k-1)}{G(k)} \geq \frac{F(k-1)}{F(k)} \\
&\iff 1 - \frac{G(k-1)}{G(k)} \leq 1 - \frac{F(k-1)}{F(k)} \\
&\iff \frac{G(k) - G(k-1)}{G(k)} \leq \frac{F(k) - F(k-1)}{F(k)} \\
&\iff \frac{g(k)}{G(k)} \leq \frac{f(k)}{F(k)}.
\end{aligned}
$$

In the above, we have defined $0/0 = 0$ wherever it may appear. Note that this does not affect the relations above since $F(k)/G(k) = 0/0$ immediately implies that $F(k-1)/G(k-1) = f(k)/F(k) = g(k)/G(k) = 0/0$. Other divide by 0 issues can be ruled out since taking $\gamma = n$, the first condition implies $F(c) \leq G(c)$ for all $c \in [n]$. $\qquad \square$

Lemma F.6 (Truncated binomial stochastic dominance). *Let $Z \sim Binom(n, p_1)$ and $Z' \sim Binom(n, p_2)$ with $p_1 \leq p_2$. Then for all c, γ, we have*

$$\mathbb{P}(Z \leq c \mid Z \leq \gamma) \geq \mathbb{P}(Z' \leq c \mid Z' \leq \gamma).$$

The inequality remains true if we replace c by a random variable C that is independent of Z and Z'.

Proof of Lemma F.6. First consider non-random c. Both sides of the inequality equal 1 when $c > \gamma$, so suppose that $c \leq \gamma$. To make use of Lemma F.5, we must verify that for all k,

$$\frac{\binom{n}{k}p_1^k(1-p_1)^{n-k}}{\sum_{j=0}^{k} \binom{n}{j}p_1^j(1-p_1)^{n-j}} \leq \frac{\binom{n}{k}p_2^k(1-p_2)^{n-k}}{\sum_{j=0}^{k} \binom{n}{j}p_2^j(1-p_2)^{n-j}}.$$

It suffices to show that for any fixed k, the left-hand side is non-decreasing in p_1. Writing $p := p_1$, we can rewrite the left-hand side as

$$\frac{\binom{n}{k}p^k(1-p)^{n-k}}{\sum_{j=0}^{k} \binom{n}{j}p^j(1-p)^{n-j}} = \left(\sum_{j=0}^{k} a_j p^{j-k}(1-p)^{n-j-(n-k)} \right)^{-1} = \left(\sum_{j=0}^{k} a_j \left(\frac{p}{1-p} \right)^{j-k} \right)^{-1}$$

where $a_j = \binom{n}{j}/\binom{n}{k}$. Since $p \to \frac{p}{1-p}$ is increasing, $j - k \leq 0$, and $a_j \geq 0$, we have that $p \to a_j \left(\frac{p}{1-p} \right)^{j-k}$ is non-increasing. The sum of non-increasing functions is non-increasing, so the sum inside the parentheses is non-increasing and positive, and therefore its inverse is non-decreasing, as desired. Hence, Lemma F.5 implies that

$$\frac{\mathbb{P}(Z' \leq c)}{\mathbb{P}(Z' \leq \gamma)} \leq \frac{\mathbb{P}(Z \leq c)}{\mathbb{P}(Z \leq \gamma)} \iff \mathbb{P}(Z' \leq c \mid Z' \leq \gamma) \leq \mathbb{P}(Z \leq c \mid Z \leq \gamma)$$

for $c \leq \gamma$. Now consider random C and let γ be fixed. Define $h(c) = \mathbb{P}(Z \leq c \mid Z \leq \gamma)$ and $h'(c) = \mathbb{P}(Z' \leq c \mid Z' \leq \gamma)$. Since C is independent of Z and Z', we have that

$$\mathbb{P}(Z \leq C \mid Z \leq \gamma) = \mathbb{E}[h(C)],$$
$$\mathbb{P}(Z' \leq C \mid Z' \leq \gamma) = \mathbb{E}[h'(C)].$$

But h dominates h' pointwise, so $h(C) \geq h'(C)$ almost surely, and $\mathbb{E}[h(C)] \geq \mathbb{E}[h'(C)]$. $\qquad \square$

Lemma F.7 (Continuous conditional stochastic dominance). *Let X, Y be continuous random variables supported on all of $\mathbb{R}$. Denote the PMF and CDF of X by f and F, and the PMF and CDF of Y by g and G, respectively. Then*

$$\frac{F(c)}{F(\gamma)} \leq \frac{G(c)}{G(\gamma)} \quad \text{for all } c \leq \gamma \in \mathbb{R} \iff \frac{f(x)}{F(x)} \geq \frac{g(x)}{G(x)} \quad \text{for all } x \in \mathbb{R}.$$

Proof of Lemma F.7. This is a standard result, but we replicate the proof here for completeness. Since $\frac{F(c)}{F(\gamma)} \leq \frac{G(c)}{G(\gamma)} \iff \frac{F(c)}{G(c)} \leq \frac{F(\gamma)}{G(\gamma)}$, the first statement is equivalent to the claim that $F(x)/G(x)$ is increasing in x, or equivalently that $\log F(x) - \log G(x)$ is increasing in x. But

$$\frac{d}{dx}\left(\log F(x) - \log G(x)\right) = \frac{f(x)}{F(x)} - \frac{g(x)}{G(x)}$$

which is greater than or equal to 0 for all x if and only if $\frac{f(x)}{F(x)} \geq \frac{g(x)}{G(x)}$ for all x. $\qquad \square$

Lemma F.8 (Truncated normal stochastic dominance). *Let $Z \sim \mathcal{N}(\mu_1, 1)$ and $Z' \sim \mathcal{N}(\mu_2, 1)$ with $\mu_1 \leq \mu_2$. Then for all c, γ, we have*

$$\mathbb{P}(Z \leq c \mid Z \leq \gamma) \geq \mathbb{P}(Z' \leq c \mid Z' \leq \gamma).$$

Letting $\Phi(c; \gamma) = \mathbb{P}(Z \leq c \mid Z \leq \gamma)$, this is equivalently stated as $\Phi(c + s; \gamma + s) \geq \Phi(c; \gamma)$ for all $c, \gamma \in \mathbb{R}$ and $s \geq 0$.

Proof of Lemma F.8. Both sides of the inequality equal 1 if $c > \gamma$, so suppose that $c \leq \gamma$. To make use of Lemma F.7, we must show that for all x,

$$\frac{\phi(x - \mu_1)}{\Phi(x - \mu_1)} \leq \frac{\phi(x - \mu_2)}{\Phi(x - \mu_2)}.$$

Since $x - \mu_1 \geq x - \mu_2$, it suffices to show that $\phi(y)/\Phi(y)$ is decreasing in y. A well-known fact is that $\phi(y)/\Phi(y)$, known as the inverse Mill's ratio, is decreasing in y. Hence it follows from Lemma F.7 that

$$\frac{\mathbb{P}(Z' \leq c)}{\mathbb{P}(Z' \leq \gamma)} \leq \frac{\mathbb{P}(Z \leq c)}{\mathbb{P}(Z \leq \gamma)} \iff \mathbb{P}(Z' \leq c \mid Z' \leq \gamma) \leq \mathbb{P}(Z \leq c \mid Z \leq \gamma)$$

for $c \leq \gamma$. $\qquad \square$

Lemma F.9 (Convergence of Kolmogorov distance of unconditional normal implies convergence of truncated variant). *Let $Z_t^{(n)}$ be a doubly-indexed sequence of random variables where t takes values in a non-random set S_n. Similarly, let $\mathcal{F}_t^{(n)}$ be a doubly-indexed sequence of σ-algebras. For each $t \in S_n$, define*

$$F_t^{(n)}(x) := \mathbb{P}(Z_t^{(n)} \leq x \mid \mathcal{F}_t^{(n)}),$$
$$F_t^{(n)}(x; y) := \mathbb{P}(Z_t^{(n)} \leq x \mid \mathcal{F}_t^{(n)}, Z_t^{(n)} \leq y).$$

Write Φ for the standard normal CDF and $\Phi(x; y) = \mathbb{P}(Z \leq x \mid Z \leq y)$ where Z is standard normal. Let A_n be a random subset of S_n for each n. Then

$$\sup_{t \in A_n} \sup_{x \in \mathbb{R}} |F_t^{(n)}(x) - \Phi(x)| \xrightarrow{p} 0 \implies \sup_{t \in A_n} \sup_{x \geq 0, y \in \mathbb{R}} |F_t^{(n)}(x; y) - \Phi(x; y)| \xrightarrow{p} 0.$$

Proof of Lemma F.9. First note that

$$\sup_{t \in A_n} \sup_{x \geq 0, y \in \mathbb{R}} |F_t^{(n)}(x; y) - \Phi(x; y)| = \sup_{t \in A_n} \sup_{0 \leq x \leq y} |F_t^{(n)}(x; y) - \Phi(x; y)|$$

since when $y < x$, both $F_t^{(n)}(x; y)$ and $\Phi(x; y)$ evaluate to 1, and hence their difference is 0. We would like to rewrite $F_t^{(n)}(x; y)$ as the ratio $F_t^{(n)}(x)/F_t^{(n)}(y)$, but this is only well-defined when the denominator is nonzero. Thus, let us first show that the denominator is nonzero with probability tending to 1:

$$
\begin{aligned}
\inf_{t\in A_n}\inf_{y\geq 0} F_t^{(n)}(y) &\geq \inf_{t\in A_n}\inf_{y\geq 0}\Phi(y) - |F_t^{(n)}(y) - \Phi(y)| \\
&\geq \inf_{t\in A_n}\inf_{y\geq 0}\Phi(y) + \inf_{t\in A_n}\inf_{y\geq 0}(-|F_t^{(n)}(y) - \Phi(y)|) \\
&\geq 0.5 - \sup_{t\in A_n}\sup_{y\geq 0}|F_t^{(n)}(y) - \Phi(y)| \\
&\xrightarrow{p} 0.5.
\end{aligned}
$$

Let $\delta, \epsilon > 0$ and choose $0 < s_1 < 0.5$ such that $\mathbb{P}(\inf_{t\in A_n}\inf_{y\geq 0} F_t^{(n)}(y) < 0.5 - s_1) < \epsilon/2$ for all sufficiently large n. Also, for notational convenience define the event $E_n := \{\inf_{t\in A_n}\inf_{y\geq 0} F_t^{(n)}(y) < 0.5 - s_1\}$.

Next, we observe that $h(x, y) = x/y$ is uniformly continuous on $[0.5 - s_1, 1]^2$. This follows since it is a continuous function on a compact set. Hence, letting $\|\cdot\|_2$ denote the standard Euclidean norm, there exists some $s_2 > 0$ such that $\|(x_1, y_1) - (x_2, y_2)\|_2 < s_2 \implies |h(x_1, y_1) - h(x_2, y_2)| < \delta$ for all $(x_1, y_1), (x_2, y_2) \in [0.5 - s_1, 1]^2$. Then for all sufficiently large n,

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{t\in A_n}\sup_{0\leq x\leq y}|F_t^{(n)}(x; y) - \Phi(x; y)| > \delta\right) \\
\leq\,&\mathbb{P}\left(\sup_{t\in A_n}\sup_{0\leq x\leq y}|F_t^{(n)}(x; y) - \Phi(x; y)| > \delta, E_n^{\mathsf{c}}\right) + \mathbb{P}(E_n) \\
=\,&\mathbb{P}\left(\sup_{t\in A_n}\sup_{0\leq x\leq y}\left|\frac{F_t^{(n)}(x)}{F_t^{(n)}(y)} - \frac{\Phi(x)}{\Phi(y)}\right| > \delta, E_n^{\mathsf{c}}\right) + \mathbb{P}(E_n) \\
=\,&\mathbb{P}\left(\sup_{t\in A_n}\sup_{0\leq x\leq y}|h(F_t^{(n)}(x), F_t^{(n)}(y)) - h(\Phi(x), \Phi(y))| > \delta, E_n^{\mathsf{c}}\right) + \mathbb{P}(E_n) \\
\leq\,&\mathbb{P}\left(\sup_{t\in A_n}\sup_{0\leq x\leq y}\|(F_t^{(n)}(x), F_t^{(n)}(y)) - (\Phi(x), \Phi(y))\|_2 \geq s_2, E_n^{\mathsf{c}}\right) + \mathbb{P}(E_n) \\
\leq\,&\mathbb{P}\left(\sup_{t\in A_n}\sup_{0\leq x\leq y}\|(F_t^{(n)}(x), F_t^{(n)}(y)) - (\Phi(x), \Phi(y))\|_2 \geq s_2\right) + \epsilon/2 \\
&\xrightarrow{p}\epsilon/2.
\end{aligned}
$$

In particular, the probability is bounded below ϵ for all sufficiently large n, as desired. The third line follows from the fact that $x \leq y$ and the denominator term $F_t^{(n)}(y)$ is positive for all $t \in A_n$ and $y \geq 0$ on the event E_n^{c}. The fifth line follows from the uniform continuity, where we additionally note that $F_t^{(n)}(x), F_t^{(n)}(y), \Phi(x), \Phi(y)$ are in $[0.5 - s_1, 1]$ for all $t \in A_n$ and $0 \leq x \leq y$ on the event E_n^{c}. The last line follows from the fact that $\sqrt{a^2 + b^2} \leq |a| + |b|$

and hence

$$\sup_{t \in A_n} \sup_{0 \leq x \leq y} \|(F_t^{(n)}(x), F_t^{(n)}(y)) - (\Phi(x), \Phi(y))\|_2$$

$$\leq \sup_{t \in A_n} \sup_{0 \leq x \leq y} |F_t^{(n)}(x) - \Phi(x)| + |F_t^{(n)}(y) - \Phi(y)|$$

$$\leq \sup_{t \in A_n} \sup_{x \in \mathbb{R}} |F_t^{(n)}(x) - \Phi(x)| + \sup_{t \in A_n} \sup_{y \in \mathbb{R}} |F_t^{(n)}(y) - \Phi(y)|$$

$$\xrightarrow{p} 0$$

by assumption. $\hspace{1cm}\square$

Lemma F.10 (Representation for uniform order statistics). *Let $U_{m,n}$ be the mth order statistic of n uniforms. Suppose that $U_{1,1}, U_{2,2}, ..., U_{n,n}$ are independent. For $i = 1, ..., n$ define*

$$Z_i = U_{n,n} \times ... \times U_{n+1-i,n+1-i}.$$

Then $(Z_n, ..., Z_1)$ has the same distribution as $(U_{(1)}, ..., U_{(n)})$, the joint distribution of the order statistics of n standard uniforms.

Proof of Lemma F.10. We proceed by induction. First, $Z_1 \overset{d}{=} U_{(n)}$ by definition. Now suppose

$$(Z_1, ..., Z_i) \overset{d}{=} (U_{(n)}, ..., U_{(n+1-i)}).$$

If $(U_1, ..., U_n)$ are standard uniform and we condition on the top i values, the remaining $n - i$ coordinates are independently distributed from $\text{Unif}[0, U_{(n+1-i)}]$, where we note that $U_{(n+1-i)}$ is the smallest of the top i values. Then the maximum among the remaining $n - i$ coordinates has the distribution $U_{(n+1-i)} \times U_{n-i,n-i}$. In other words, the distribution of $U_{(n-i)}$ conditional on $(U_{(n)}, ..., U_{(n+1-i)})$ is $U_{(n+1-i)} \times U_{n-i,n-i}$. Similarly, conditional on $(Z_1, ..., Z_i)$, the distribution of Z_{i+1} is $Z_i \times U_{n-i,n-i}$. Thus, applying the inductive hypothesis we have that $(Z_1, ..., Z_{i+1}) \overset{d}{=} (U_{(n)}, ..., U_{(n-i)})$. $\hspace{1cm}\square$

Lemma F.11 (Positive region conditional variance). *Suppose $X \in \mathcal{X}$ and $Y \in \mathbb{R}$ are such that $\text{Var}(Y \mid X) > 0$ almost surely. Assume that all $R \subseteq \mathcal{X}$ referred to are measurable. For $R \subseteq \mathcal{X}$, define $\rho(R) := \mathbb{P}(X \in R)$. Then*

$$\inf_{R \subseteq \mathcal{X}: \rho(R) \geq q} \text{Var}(Y \mid X \in R) > 0$$

for all $0 < q \leq 1$.

Proof of Lemma F.11. First observe that $\text{Var}(Y \mid X \in R) \geq \mathbb{E}[\text{Var}(Y \mid X) \mid X \in R]$ by the law of total variance. Let us simply denote $f(X) := \text{Var}(Y \mid X)$ for notational convenience. It suffices to show that

$$\inf_{R \subseteq \mathcal{X}: \rho(R) \geq q} \mathbb{E}[f(X) \mid X \in R] > 0.$$

121

Suppose for contradiction that the infimum is 0. Then there exists a sequence $(R_n)_{n=1}^\infty$ such that $\mathbb{E}[f(X) \mid X \in R_n] \leq n^{-2}$ and $\rho(R_n) \geq q$ for all n. Then

$$
\begin{aligned}
&\mathbb{P}(f(X) > n^{-1}) \\
=&\mathbb{P}(f(X) > n^{-1} \mid X \in R_n)\mathbb{P}(X \in R_n) + \mathbb{P}(f(X) > n^{-1} \mid X \notin R_n)\mathbb{P}(X \notin R_n) \\
\leq&n \cdot \mathbb{E}[f(X) \mid X \in R_n]\mathbb{P}(X \in R_n) + \mathbb{P}(X \notin R_n) \\
\leq&n \cdot \frac{1}{n^2} \cdot \mathbb{P}(X \in R_n) + \mathbb{P}(X \notin R_n) \\
\leq&\frac{1}{n} \cdot q + (1-q).
\end{aligned}
$$

In going to the third line of the above, we use the fact that $\mathbb{P}(f(X) > n^{-1} \mid X \in R_n) \leq \mathbb{P}(f(X) \geq n^{-1} \mid X \in R_n)$, and then we apply Markov's inequality conditionally on the event $\{X \in R_n\}$. Next, by the right continuity of CDFs, for all $\epsilon > 0$ there exists an N such that for all $n \geq N$, $|\mathbb{P}(f(X) \leq 0) - \mathbb{P}(f(X) \leq n^{-1})| \leq \epsilon$. Then

$$
\begin{aligned}
\mathbb{P}(f(X) = 0) = \mathbb{P}(f(X) \leq 0) \\
\geq \mathbb{P}(f(X) \leq n^{-1}) - \epsilon \\
\geq 1 - \frac{1}{n} \cdot q - (1-q) - \epsilon \\
= q\left(1 - \frac{1}{n}\right) - \epsilon
\end{aligned}
$$

for all $n \geq N$. In particular, choosing $\epsilon = q/2$, for sufficiently large n the above implies that $\mathbb{P}(f(X) = 0)$ is positive, which contradicts our assumption that $\mathrm{Var}(Y \mid X) > 0$ almost surely. Hence, the infimum cannot be 0 and must be positive. $\qquad\square$

Lemma F.12 (Simple centered moment bound using raw moments). *Let k be a fixed positive integer greater than 1 and let Y be a random variable with finite kth moment. Then*

$$
\mathbb{E}[|Y - \mathbb{E}[Y]|^k] \leq 2^k \cdot \mathbb{E}[|Y|^k].
$$

Proof of Lemma F.12. It is a well-known fact that for random variables X_1, X_2 such that the kth moment of each exists, we have

$$
\mathbb{E}[|X_1 + X_2|^k] \leq 2^{k-1}(\mathbb{E}[|X_1|^k] + \mathbb{E}[|X_2|^k]).
$$

See, for instance, the introduction of Bahr and Esseen (1965). Letting $X_1 = Y$ and $X_2 = -\mathbb{E}[Y]$, the above yields

$$
\begin{aligned}
\mathbb{E}[|Y - \mathbb{E}[Y]|^k] &\leq 2^{k-1}(\mathbb{E}[|Y|^k] + \mathbb{E}[|\mathbb{E}[Y]|^k]) \\
&= 2^{k-1}(\mathbb{E}[|Y|^k] + |\mathbb{E}[Y]|^k) \\
&\leq 2^{k-1}(\mathbb{E}[|Y|^k] + \mathbb{E}[|Y|^k]) \\
&= 2^k \cdot \mathbb{E}[|Y|^k]
\end{aligned}
$$

where we use the fact that $|\mathbb{E}[Y]|^k \leq \mathbb{E}[|Y|^k]$ by Jensen's inequality. $\qquad\square$

Lemma F.13 (Conditional vanishing in probability). *Let X_n be a sequence of random variables and $\mathcal{F}_n$ be a sequence of σ-algebras. Suppose that $\mathbb{E}[X_n \mid \mathcal{F}_n] = 0$ and $\mathrm{Var}(X_n \mid \mathcal{F}_n) \xrightarrow{p} 0$. Then $X_n \xrightarrow{p} 0$.*

Proof of Lemma F.13. Define $\sigma_n^2 := \mathrm{Var}(X_n \mid \mathcal{F}_n)$, which is a random variable. For any $\delta, \epsilon > 0$, let $b := \epsilon \delta^2 / 2$ and choose N so that $\mathbb{P}(\sigma_n^2 > b) < \epsilon/2$ for all $n \geq N$. Then since σ_n^2 is $\mathcal{F}_n$-measurable,

$$\mathbb{P}(|X_n| > \delta, \sigma_n^2 \leq b \mid \mathcal{F}_n) = \mathbb{1}\{\sigma_n^2 \leq b\}\mathbb{P}(|X_n| > \delta \mid \mathcal{F}_n) \leq \mathbb{1}\{\sigma_n^2 \leq b\} \cdot \frac{\sigma_n^2}{\delta^2} \leq b/\delta^2 = \epsilon/2$$

by conditional Chebyshev's inequality. Then taking expectations of both sides, $\mathbb{P}(|X_n| > \delta, \sigma_n^2 \leq b) \leq \epsilon/2$. Then

$$\begin{aligned}
\mathbb{P}(|X_n| > \delta) &\leq \mathbb{P}(|X_n| > \delta, \sigma_n^2 \leq b) + \mathbb{P}(\sigma_n^2 > b) \\
&\leq \epsilon/2 + \epsilon/2 \\
&= \epsilon
\end{aligned}$$

for all $n \geq N$. $\qquad\square$

F.4 Convergence on bounded metrics

A difficulty that we encounter throughout is that certain quantities do not converge in the usual sense, but are close in the sense that they are either close to one another, or they diverge together in the same direction. For our purposes, this weaker notion of convergence suffices. To formalize this notion, we introduce a new metric under which such quantities *do* converge, and establish some useful properties. Throughout this section, f will always refer to a strictly increasing function. Occasionally, we will require f to have additional properties, but we will always specify these situations.

Definition F.14 (f-metric). *Let $f : \mathbb{R} \to \mathbb{R}$ be a strictly increasing function, and by an abuse of notation let $f(x) = (f(x_1), ..., f(x_k))$ for $x \in \mathbb{R}^k$. We define $d_f : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ via*

$$d_f(x, y) = \|f(x) - f(y)\|_2$$

where $\|\cdot\|_2$ is the standard Euclidean norm. We will refer to d_f as the f-metric.

Lemma F.15 (Metricity). *d_f is a metric.*

Proof of Lemma F.15. Symmetry and the fact that $d_f(x, x) = 0$ are immediate. Positivity follows from the fact that f is strictly increasing and so $f(x) \neq f(y)$ for all $x \neq y$. Lastly,

$$\begin{aligned}
d_f(x, z) &= \|f(x) - f(z)\|_2 \\
&= \|f(x) - f(y) + f(y) - f(z)\|_2 \\
&\leq \|f(x) - f(y)\|_2 + \|f(y) - f(z)\|_2 \\
&= d_f(x, y) + d_f(x, z)
\end{aligned}$$

from the ordinary triangle inequality on $\|\cdot\|_2$. $\qquad\square$

We extend the notion of convergence in probability to f-metrics.

Definition F.16 (f-convergence). *Let X_n and Y_n be two sequences of random variables. We will say that X_n converges to Y_n under the f-metric if $d_f(X_n, Y_n) \xrightarrow{p} 0$. In this case, we will write $X_n \overset{f}{\leftrightarrow} Y_n$, with the bidirectional arrow to indicate that the relation is symmetric; that is, $X_n \overset{f}{\leftrightarrow} Y_n \iff Y_n \overset{f}{\leftrightarrow} X_n$.*

We will be able to show that f-convergence is preserved under uniformly continuous maps, where uniform continuity is defined with respect to the f-metric. However, some maps that we need are not uniformly continuous maps. But as long as the random variables that are inputs to this map "usually" lie on a uniformly continuous restriction, we will face no issues. To this end, we introduce the following notion, which captures the usual notion of bounded in probability, but also handles the idea of a sequence of random variables being "bounded away" from some threshold. We also show that this notion of boundedness is preserved between f-convergent sequences.

Definition F.17 ($\mathcal{B}$-bounded in probability). *Let X_n be a sequence of random variables, and let $\mathcal{B}$ be a collection of open intervals in $\mathbb{R}$ with the property that, for every $(a, b) \in \mathcal{B}$, there exists $(a', b') \in \mathcal{B}$ such that $a' < a$ and $b' > b$, where we take the convention that $-\infty < -\infty$ and $\infty > \infty$. We say that X_n is $\mathcal{B}$-bounded in probability if for every $\epsilon > 0$, there exists an integer N and an interval $I \in \mathcal{B}$ such that $\mathbb{P}(X_n \notin I) < \epsilon$ for all $n > N$. We note three special cases:*

1. *When $\mathcal{B} = \{(-c, c) : c > 0\}$, we will simply say that X_n is bounded in probability.*

2. *When $\mathcal{B} = \{(1/c, c) : c > 1\}$, we will simply say that X_n is positively log-bounded in probability.*

3. *When $(-\infty, \infty) \in \mathcal{B}$, then every X_n is trivially $\mathcal{B}$-bounded in probability.*

Lemma F.18 (Transitivity of $\mathcal{B}$-boundedness in probability). *Let X_n and Y_n be sequences of random variables and suppose $X_n \overset{f}{\leftrightarrow} Y_n$. Let $\mathcal{B}$ be any collection that satisfies the conditions in Definition F.17. Then X_n is $\mathcal{B}$-bounded in probability if and only if Y_n is $\mathcal{B}$-bounded in probability.*

Proof of Lemma F.18. Without loss of generality suppose X_n is $\mathcal{B}$-bounded in probability. Let $\epsilon > 0$. Choose $I_1 \in \mathcal{B}$ and N_1 such that $\mathbb{P}(X_n \notin I_1) < \epsilon/2$ for all $n > N_1$. Writing $I_1 = (l_1, u_1)$ we have that by assumption there exists $I_2 = (l_2, u_2) \in \mathcal{B}$ with $l_2 < l_1$ and $u_2 > u_1$. We will just discuss the case when $l_1 \neq -\infty$ and $u_1 \neq \infty$; when either of the endpoints is infinite, the proof is very similar.

Let $\Delta = \min\{f(u_2) - f(u_1), f(l_1) - f(l_2)\}$ and note that $\Delta > 0$ since f is strictly increasing. Also note that

$$Y_n \notin I_2 \text{ and } X_n \in I_1 \implies f(Y_n) - f(X_n) > \Delta$$

because either Y_n falls to the right of u_2, in which case $|f(Y_n) - f(X_n)| > f(u_2) - f(u_1)$, or it falls to the left of l_2, in which case $|f(Y_n) - f(X_n)| > f(l_1) - f(l_2)$. Thus, the absolute

difference is always at least as great as the minimum of these two quantities. Since $X_n \overset{f}{\leftrightarrow} Y_n$, we can choose N_2 such that $\mathbb{P}(|f(X_n) - f(Y_n)| > \Delta) < \epsilon/2$ for all $n > N_2$. Then for all $n > \max\{N_1, N_2\}$,

$$
\begin{aligned}
\mathbb{P}(Y_n \notin I_2) &\leq \mathbb{P}(Y_n \notin I_2 \text{ and } X_n \in I_1) + \mathbb{P}(X_n \notin I_1) \\
&\leq \mathbb{P}(|f(Y_n) - f(X_n)| > \Delta) + \epsilon/2 \\
&\leq \epsilon/2 + \epsilon/2 \\
&= \epsilon.
\end{aligned}
$$

So Y_n is $\mathcal{B}$-bounded in probability. $\qquad\square$

We are ready to state the main composition results. We first establish a lemma regarding the leap from univariate f-convergence to bivariate f-convergence. We then give conditions for transformations that preserve f-convergence.

Lemma F.19 (Bivariate f-convergence). *Let A_n, B_n, C_n, D_n be sequences of random variables. Suppose that $A_n \overset{f}{\leftrightarrow} B_n$ and $C_n \overset{f}{\leftrightarrow} D_n$. Then*

$$
d_f((A_n, C_n), (B_n, D_n)) \overset{p}{\to} 0.
$$

Proof of Lemma F.19. Since $\sqrt{x^2 + y^2} \leq |x| + |y|$,

$$
\|(f(A_n), f(C_n)) - (f(B_n), f(D_n))\|_2 \leq |f(A_n) - f(B_n)| + |f(C_n) - f(D_n)| \overset{p}{\to} 0.
$$

$\qquad\square$

For two metrics d_X and d_Y on sets M_X and M_Y respectively, we will say that a function $g : M_X \to M_Y$ is (d_X, d_Y)-uniformly continuous if it is uniformly continuous with respect to the metric d_X on M_X and the metric d_Y on M_Y.

Lemma F.20 (Abstract f-convergence composition rule). *Let A_n, B_n, C_n, D_n be sequences of random variables. Suppose that $A_n \overset{f}{\leftrightarrow} B_n$ and $C_n \overset{f}{\leftrightarrow} D_n$. Suppose that B_n is $\mathcal{B}_1$-bounded in probability and D_n is $\mathcal{B}_2$-bounded in probability. Let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function, and for each $I_1 \in \mathcal{B}_1, I_2 \in \mathcal{B}_2$ define the restriction $g|_{I_1 \times I_2} : I_1 \times I_2 \to \mathbb{R}$ by $g|_{I_1 \times I_2}(x, y) = g(x, y)$ for all $x \in I_1, y \in I_2$. Suppose that $g|_{I_1 \times I_2}$ is uniformly (d_f, d_f)-continuous for every $I_1 \in \mathcal{B}_1, I_2 \in \mathcal{B}_2$. Then $g(A_n, C_n) \overset{f}{\leftrightarrow} g(B_n, D_n)$.*

Proof of Lemma F.20. Let $\epsilon, \delta > 0$. By Lemma F.18, we also have that A_n is $\mathcal{B}_1$-bounded in probability and C_n is $\mathcal{B}_2$-bounded in probability. Choose $I_1 \in \mathcal{B}_1, I_2 \in \mathcal{B}_2$ and N such that

$$
\begin{aligned}
1 - \mathbb{P}(\text{In}(A_n)) := \mathbb{P}(\text{Out}(A_n)) &:= \mathbb{P}(A_n \notin I_1) < \epsilon/8, \\
1 - \mathbb{P}(\text{In}(B_n)) := \mathbb{P}(\text{Out}(B_n)) &:= \mathbb{P}(B_n \notin I_1) < \epsilon/8, \\
1 - \mathbb{P}(\text{In}(C_n)) := \mathbb{P}(\text{Out}(C_n)) &:= \mathbb{P}(C_n \notin I_2) < \epsilon/8, \\
1 - \mathbb{P}(\text{In}(D_n)) := \mathbb{P}(\text{Out}(D_n)) &:= \mathbb{P}(D_n \notin I_2) < \epsilon/8
\end{aligned}
$$

for all $n > N$. Since $g|_{I_1 \times I_2}$ is uniformly (d_f, d_f)-continuous by assumption, choose $s > 0$ so that $d_f((x_1, y_1), (x_2, y_2)) < s \implies d_f(g|_{I_1 \times I_2}(x_1, y_1), g|_{I_1 \times I_2}(x_2, y_2)) < \delta$ for every $(x_1, y_1), (x_2, y_2) \in I_1 \times I_2$. Then for all $n > N$, we have

$$\mathbb{P}(d_f(g(A_n, C_n), g(B_n, D_n)) > \delta)$$
$$\leq \mathbb{P}(d_f(g(A_n, C_n), g(B_n, D_n)) > \delta, \text{In}(A_n), \text{In}(B_n), \text{In}(C_n), \text{In}(D_n))$$
$$\quad + \mathbb{P}(\text{Out}(A_n)) + \mathbb{P}(\text{Out}(B_n)) + \mathbb{P}(\text{Out}(C_n)) + \mathbb{P}(\text{Out}(D_n))$$
$$\leq \mathbb{P}(d_f(g_{I_1 \times I_2}(A_n, C_n), g_{I_1 \times I_2}(B_n, D_n)) > \delta, \text{In}(A_n), \text{In}(B_n), \text{In}(C_n), \text{In}(D_n))$$
$$\quad + 4(\epsilon/8)$$
$$\leq \mathbb{P}(d_f((A_n, C_n), (B_n, D_n)) \geq s) + \epsilon/2.$$

By Lemma F.19, $d_f((A_n, C_n), (B_n, D_n)) \xrightarrow{p} 0$, so that we can choose N' such that

$$\mathbb{P}(d_f((A_n, C_n), (B_n, D_n)) \geq s) < \epsilon/2$$

for all $n > N'$. Then $\mathbb{P}(d_f(g(A_n, C_n), g(B_n, D_n)) > \delta) < \epsilon$ for all $n > \max\{N, N'\}$. $\qquad\square$

Lemma F.20 gives us abstract conditions on maps that preserve f-convergence. The following lemmas give us a concrete strategy for establishing f-convergence. First we establish a reduction to uniform continuity with respect to the standard Euclidean metric.

Lemma F.21 (Reduction to standard uniform continuity). *Let f_1 and f_2 be strictly increasing functions from $\mathbb{R}$ to $\mathbb{R}$, and let $g : S \to \mathbb{R}$ where $S \subseteq \mathbb{R}^k$. Recall the abuse of notation that $f(x) = (f(x_1), ..., f(x_k))$ for $x \in \mathbb{R}^k$. Further abusing notation, write $f^{-1}(x) = (f^{-1}(x_1), ..., f^{-1}(x_k))$ for $x \in \mathbb{R}^k$. Then g is uniformly (d_{f_1}, d_{f_2})-continuous if and only if the function*

$$h : f_1(S) \to f_2(\mathbb{R}),$$
$$h(x) = f_2(g(f_1^{-1}(x)))$$

is uniformly continuous with respect to the standard Euclidean metric.

Proof of Lemma F.21. g is uniformly (d_{f_1}, d_{f_2})-continuous if and only if

$$\forall \epsilon \, \exists \delta \text{ s.t. } \|f_1(x) - f_1(y)\|_2 \geq \delta \text{ or } \|f_2(g(x)) - f_2(g(y))\|_2 < \epsilon \text{ for all } x, y \in S.$$

Rewriting $x' = f_1(x), y' = f_1(y)$ and noting that f_1 is invertible, the above is equivalent to

$$\forall \epsilon \, \exists \delta \text{ s.t. } \|x' - y'\|_2 \geq \delta \text{ or } \|f_2(g(f_1^{-1}(x'))) - f_2(g(f_1^{-1}(y')))\|_2 < \epsilon \text{ for all } x', y' \in f_1(S),$$

which is equivalent to h being uniformly continuous on the standard Euclidean metric. $\qquad\square$

Next, we recall the definition and some properties of uniformly equivalent metrics.

Definition F.22 (Uniformly equivalent metrics). *Let d_1 and d_2 be two metrics defined on a set M. Define $Id : (M, d_1) \to (M, d_2)$ via $Id(x) = x$. Then we will say that d_1 and d_2 are uniformly equivalent if Id and Id^{-1} are uniformly continuous.*

Lemma F.23 (Uniform equivalence of continuous bounded f-metrics)**.** *Let f_1 and f_2 be continuous, strictly increasing, and bounded functions. Then d_{f_1} and d_{f_2} are uniformly equivalent.*

Proof of Lemma F.23. Note that $f_2 \circ \mathrm{Id} \circ f_1^{-1} = f_2 \circ f_1^{-1}$. By Lemma F.21, to show that Id is uniformly continuous, it suffices to show that $f_2 \circ f_1^{-1}$, as a function from $f_1(\mathbb{R})^k$ to $f_2(\mathbb{R})^k$, is uniformly continuous with respect to the standard Euclidean metric. Recall that for $x \in f_1(\mathbb{R})^k$,

$$f_2(f_1^{-1}(x)) = (f_2(f_1^{-1}(x_1)), ..., f_2(f_1^{-1}(x_k))).$$

Let h denote $f_2 \circ f_1^{-1}$ as a univariate function from $\mathbb{R}$ to $\mathbb{R}$, i.e. $h(x) = f_2(f_1^{-1}(x))$ for $x \in f_1(\mathbb{R})$. Then h is continuous, bounded (since f_1 is bounded), and strictly increasing since both f_1^{-1} and f_2 are strictly increasing. A standard fact from real analysis is that continuous, bounded, and strictly increasing functions are uniformly continuous, so h is uniformly continuous. Furthermore, the concatenation of uniformly continuous functions is uniformly continuous, so $f_2 \circ f_1^{-1}$ (as a function from $f_1(\mathbb{R})^k$ to $f_2(\mathbb{R})^k$) is indeed uniformly continuous. An analogous argument shows that $f_1 \circ f_2^{-1}$ is uniformly continuous and hence Id^{-1} is uniformly continuous. $\square$

Lemma F.24 (Uniform equivalence preserves uniform continuity)**.** *Let M_X, M_Y be sets and suppose that d_1^X, d_2^X are uniformly equivalent metrics on M_X and d_1^Y, d_2^Y are uniformly equivalent metrics on M_Y. Let g be a function from $M_X \to M_Y$. Then g is uniformly (d_1^X, d_1^Y)-continuous if and only if g is uniformly (d_2^X, d_2^Y)-continuous.*

Proof of Lemma F.24. Suppose g is uniformly (d_1^X, d_1^Y)-continuous. Let $\mathrm{Id}_X : (M_X, d_1^X) \to (M_X, d_2^X)$ and $\mathrm{Id}_Y : (M_Y, d_1^Y) \to (M_Y, d_2^Y)$ be defined by $\mathrm{Id}_X(x) = x$ and $\mathrm{Id}_Y(y) = y$. Then note that $\mathrm{Id}_Y \circ g \circ \mathrm{Id}_X^{-1}$ is uniformly (d_2^X, d_2^Y)-continuous since g, Id_X^{-1}, and Id_Y are uniformly continuous on their respective metrics by assumption, and the composition of uniformly continuous functions is uniformly continuous. But as a function, $\mathrm{Id}_Y \circ g \circ \mathrm{Id}_X^{-1}$ is just g, so g is uniformly (d_2^X, d_2^Y)-continuous. The proof of the converse statement is identical. $\square$

With these tools, we can finally establish that certain useful transformations preserve f-convergence within the class of continuous, bounded f. Note that in particular Φ, the standard normal CDF, belongs to this class.

Lemma F.25 (Concrete f-convergence composition rules)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous, strictly increasing, and bounded function. Let A_n, B_n, C_n, D_n be sequences of random variables and suppose that $A_n \overset{f}{\leftrightarrow} B_n$ and $C_n \overset{f}{\leftrightarrow} D_n$. We have the following composition rules:*

1. *(Addition) $A_n + C_n \overset{f}{\leftrightarrow} B_n + D_n$ if B_n is bounded in probability.*

2. *(Subtraction) $A_n - C_n \overset{f}{\leftrightarrow} B_n - D_n$ if B_n is bounded in probability.*

3. *(Multiplication) $A_n C_n \overset{f}{\leftrightarrow} B_n D_n$ if B_n is positively log-bounded in probability.*

4. *(Minimum) $\min\{A_n, C_n\} \overset{f}{\leftrightarrow} \min\{B_n, D_n\}$.*

5. *(Clipped truncated normal quantile) Define* $\mathrm{Clip}_{a,b}(x) = \max\{a, \min\{x, b\}\}$. *For any fixed* $b \in (0, 1)$*, let*

$$g(x, y) = \mathrm{Clip}_{0,\infty}\big(\Phi^{-1}(\mathrm{Clip}_{0,b}(x) \cdot \Phi(y))\big).$$

Then $g(A_n, C_n) \overset{f}{\leftrightarrow} g(B_n, D_n)$.

Proof of Lemma F.25. Define $f(x) = \frac{1}{1 + e^{-x}}$ and note that $f^{-1}(x) = \log \frac{x}{1-x}$. Since f is continuous, strictly increasing, and bounded, by Lemmas F.23 and F.24, it suffices to establish the composition rules for this specific choice of f. In particular, we will establish the condition in Lemma F.21 for this choice of f, and then apply Lemma F.20.

(Addition) Let $g(x, y) = x + y$. Note that

$$g(f^{-1}(x), f^{-1}(y)) = \log \frac{x}{1 - x} + \log \frac{y}{1 - y}$$

$$\implies \exp\big(g(f^{-1}(x), f^{-1}(y))\big) = \frac{x}{1 - x} \frac{y}{1 - y}$$

$$\implies h(x, y) := f(g(f^{-1}(x), f^{-1}(y))) = \frac{1}{1 + \left(\frac{x}{1-x}\frac{y}{1-y}\right)^{-1}} = \frac{xy}{1 - x - y + 2xy}.$$

Consider the restriction of g to $(-c, c) \times \mathbb{R}$ for any $c > 0$. Then since $f(\mathbb{R}) = (0, 1)$ and $f((-c, c)) = (a, b)$ for some $0 < a < b < 1$, Lemma F.21 implies that it suffices to show that h is uniformly continuous on $(a, b) \times (0, 1)$ for all $0 < a < b < 1$. For each fixed x, the expression $1 - x - y + 2xy$ is monotonic in y, and so the minimum is attained either when $y = 0$ or $y = 1$. Using this, it is straightforward to check that $1 - x - y + 2xy \geq \min\{a, 1 - b\} > 0$ for all $(x, y) \in [a, b] \times [0, 1]$. So we have that h is continuous, and hence uniformly continuous, on $[a, b] \times [0, 1]$. Thus it is also uniformly continuous on the restriction to $(a, b) \times (0, 1)$.

(Subtraction) Let $g(x, y) = x - y$. A similar manipulation to the previous case yields

$$h(x, y) := f(g(f^{-1}(x), f^{-1}(y))) = \frac{1}{1 + \left(\frac{x}{1-x}\frac{1-y}{y}\right)^{-1}} = \frac{x(1 - y)}{x(1 - y) + (1 - x)y}.$$

Again, for h restricted to $(a, b) \times (0, 1)$ for any $0 < a < b < 1$, the denominator is no smaller than $\min\{a, 1 - b\} > 0$, so h is continuous, and hence uniformly continuous, on $[a, b] \times [0, 1]$. Thus, h is also uniformly continuous on the restriction to $(a, b) \times (0, 1)$.

(Multiplication) Let $g(x, y) = xy$. Note that

$$g(f^{-1}(x), f^{-1}(y)) = \log \frac{x}{1 - x} \cdot \log \frac{y}{1 - y}$$

$$\implies \exp\big(g(f^{-1}(x), f^{-1}(y))\big) = \left(\frac{y}{1 - y}\right)^{\log \frac{x}{1-x}}$$

$$\implies h(x, y) := f(g(f^{-1}(x), f^{-1}(y))) = \frac{1}{1 + \left(\frac{y}{1-y}\right)^{-\log \frac{x}{1-x}}}.$$

Consider the restriction of g to $(1/c, c) \times \mathbb{R}$ for any $c > 1$. Then $f((1/c, c)) = (a, b)$ for some $0.5 < a < b < 1$. Thus, Lemma F.21 implies that it suffices to show that h is uniformly continuous on $(a, b) \times (0, 1)$ for all $0.5 < a < b < 1$. Note that

$$h_2(z, y) := \frac{1}{1 + \left(\frac{y}{1-y}\right)^{-z}} = \frac{1}{\frac{y^z}{y^z} + \frac{(1-y)^z}{y^z}} = \frac{y^z}{y^z + (1-y)^z}$$

is continuous for $(z, y) \in I \times [0, 1]$ for any compact interval I that excludes 0, since then the denominator $y^z + (1-y)^z$ is well-defined and strictly greater than 0. Furthermore, we have that $h_3(x, y) := \left(\log \frac{x}{1-x}, y\right)$ is a continuous function for $(x, y) \in [a, b] \times [0, 1]$, and moreover the range of h_3 is contained in $I \times [0, 1]$ for some compact interval I that excludes 0. Hence, $h_2 \circ h_3$ is a continuous function on $[a, b] \times [0, 1]$, and hence uniformly continuous. But $h = h_2 \circ h_3$ on the restriction to $(a, b) \times (0, 1)$, so it is uniformly continuous on $(a, b) \times (0, 1)$.

(Minimum) Let $g(x, y) = \min\{x, y\}$. Note that

$$g(f^{-1}(x), f^{-1}(y)) = \min\left\{\log \frac{x}{1 - x}, \log \frac{y}{1 - y}\right\}$$

$$\implies \exp\left(g(f^{-1}(x), f^{-1}(y))\right) = \min\left\{\frac{x}{1 - x}, \frac{y}{1 - y}\right\}$$

$$\implies h(x, y) := f(g(f^{-1}(x), f^{-1}(y))) = \frac{1}{1 + \min\left\{\frac{x}{1-x}, \frac{y}{1-y}\right\}^{-1}} = \min\{x, y\}.$$

We see that h is continuous and hence uniformly continuous on $[0, 1] \times [0, 1]$. Thus, it is uniformly continuous on the restriction to $(0, 1) \times (0, 1)$.

(Clipped truncated normal quantile) For this last rule, we will establish the conditions in Lemma F.21 with $f = \Phi$. In this case,

$$g(f^{-1}(x), f^{-1}(y)) = \text{Clip}_{0,\infty}\left(\Phi^{-1}\left(\text{Clip}_{0,b}(\Phi^{-1}(x)) \cdot y\right)\right)$$

$$\implies h(x, y) := f(g(f^{-1}(x), f^{-1}(y))) = \text{Clip}_{0.5,1}\left(\text{Clip}_{0,b}(\Phi^{-1}(x)) \cdot y\right).$$

Next note that

$$\text{Clip}_{0.5,1}\left(\text{Clip}_{0,b}(\Phi^{-1}(x)) \cdot y\right) = \text{Clip}_{0.5,1}\left(\text{Clip}_{0.5,b}(\Phi^{-1}(x)) \cdot y\right)$$

$$= \text{Clip}_{0.5,1}\left(\Phi^{-1}(\text{Clip}_{\Phi(0.5),\Phi(b)}(x)) \cdot y\right)$$

where the first equality follows from the fact that if $\Phi^{-1}(x) \leq 0.5$, then the largest value that $\text{Clip}_{0,b}(\Phi^{-1}(x)) \cdot y$ can take over $y \in [0, 1]$ is 0.5. Since this quantity gets clipped to the range $[0.5, 1]$ regardless, it makes no difference if we apply an additional clipping that affects values outside of this range prior to clipping to $[0.5, 1]$. Then note that $\text{Clip}_{\Phi(0.5),\Phi(b)}(x)$ is a continuous function on $[0, 1]$ whose range is $[\Phi(0.5), \Phi(b)]$, and that Φ^{-1} is continuous on $[\Phi(0.5), \Phi(b)]$. Hence, $h(x, y)$ is continuous, and thus uniformly continuous, on $[0, 1] \times [0, 1]$. In particular, it is uniformly continuous on the restriction to $(0, 1) \times (0, 1)$. $\square$

We state one more general composition rule with a simpler proof than those above.

Lemma F.26 (Indicator f-convergence composition rule)**.** *Suppose that $X_n \overset{f}{\leftrightarrow} Y_n$ and that A_n is a sequence of binary random variables. Then $X_n \cdot A_n \overset{f}{\leftrightarrow} Y_n \cdot A_n$.*

Proof of Lemma F.26. The result follows from

$$d_f(X_n \cdot A_n, Y_n \cdot A_n) = d_f(X_n, Y_n) \cdot A_n \leq d_f(X_n, Y_n) \overset{p}{\to} 0.$$

$\square$

Lastly, we state a result that connects f-convergence to convergence under the usual Euclidean metric for bounded sequences.

Lemma F.27 (*f-convergence to ordinary convergence*)**.** *Let f be a continuous, strictly increasing, and bounded function, and suppose $X_n \overset{f}{\leftrightarrow} Y_n$. Furthermore, suppose X_n and Y_n are bounded a.s. by some constant M for all n. Then*

$$|X_n - Y_n| \overset{p}{\to} 0.$$

Proof of Lemma F.27. Let $\epsilon, \delta > 0$. Let g denote the inverse of f, which exists and is continuous. Note that g is uniformly continuous on the compact interval $I := [f(-M), f(M)]$. Thus, choose s so that $|x - y| < s \implies |g(x) - g(y)| < \delta$ for all $x, y \in I$. By assumption, there exists an N such that $\mathbb{P}(|f(X_n) - f(Y_n)| \geq s) < \epsilon$ for all $n > N$. Then

$$\mathbb{P}(|X_n - Y_n| > \delta) = \mathbb{P}(|g(f(X_n)) - g(f(Y_n))| > \delta) \leq \mathbb{P}(|f(X_n) - f(Y_n)| \geq s) < \epsilon.$$

$\square$