Statistica Sinica Preprint No: SS-2024-0013	
Title	Estimation of Subsidiary Performance Metrics under Optimal Policies
Manuscript ID	SS-2024-0013
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0013
Complete List of Authors	Zhaoqi Li, Houssam Nassif and Alex Luedtke
Corresponding Authors	Zhaoqi Li
E-mails	zli9@stanford.edu
Notice: Accepted author version.	

Statistica Sinica: Newly accepted Paper

Statistica Sinica

Estimation of subsidiary performance metrics under optimal policies

Zhaoqi Li*, Houssam Nassif†, Alex Luedtke*

*University of Washington, †Meta

Abstract:

In policy learning, the goal is typically to optimize a primary performance metric, but other subsidiary metrics often also warrant attention. This paper presents two strategies for evaluating these subsidiary metrics under a policy that is optimal for the primary one. The first relies on a novel margin condition that facilitates Wald-type inference. Under this and other regularity conditions, we show that the one-step corrected estimator is efficient. Despite the utility of this margin condition, it places strong restrictions on how the subsidiary metric behaves for nearly-optimal policies, which may not hold in practice. We therefore introduce alternative, two-stage strategies that do not require a margin condition. The first stage constructs a set of candidate policies and the second builds a uniform confidence interval over this set. We provide numerical simulations to evaluate the performance of these methods in different scenarios.

Key words and phrases: Statistical inference, optimal policies, efficient estimator.

1. Introduction

1.1 Literature Review

Many fields are interested in learning policies that map from individual-level characteristics to a set of actions. The policies that result in the best possible mean of a subsequent outcome are often referred to as optimal policies [Athey and Wager, 2021]. For example, in biomedical sciences the action may take the form of a treatment allocation and the outcome may be disease remission [Ling et al., 2023], whereas in digital marketing the action and outcome may be a recommendation and click-through rate, respectively [Fiez et al., 2024]. Various methods have been developed to estimate optimal policies. Examples include Q-learning, which uses regression modeling of the expected outcome to guide policy decisions [Qian and Murphy, 2011], outcome-weighted learning [Zhao et al., 2012], and doubly robust methods [Murphy, 2003, Robins, 2004, Dudík et al., 2011, Zhang et al., 2013]. Performance guarantees for these methods have been established by several authors [Qian and Murphy, 2011, Zhao et al., 2012, Luedtke and Chambaz, 2020, Athey and Wager, 2021].

An estimated policy is unlikely to be implemented unless confidence intervals characterizing its performance are available [Shi et al., 2021, Weltz

Statistica Sinica: Newly accepted Paper

et al., 2023]. These performance metrics may take the form of the remission rate of all patients or the click-through rate of all customers. In both of these examples, the metric is the value of the optimal policy in the population, better known as the optimal value. Inference about the optimal value is well-studied when there is only one outcome of interest [Luedtke and Chambaz, 2020, Liu et al., 2021]. Several works have shown that one-step estimators and targeted minimum loss-based estimators are efficient under certain conditions [van der Laan and Luedtke, 2015, Chambaz et al., 2017]. In particular, these works require a non-exceptional law condition that states that the conditional average action effect does not concentrate mass at zero [Robins, 2004]. Alternative strategies have been developed for constructing confidence intervals for the optimal value even when this condition fails [Chakraborty et al., 2013, Luedtke and Van Der Laan, 2016, Shi et al., 2021].

Though most existing methodological works on policy learning focus on optimizing for a single performance metric, in real-world settings there are often multiple other subsidiary performance metrics that are also of interest [Boominathan et al., 2020, Bica et al., 2021]. These metrics may correspond to different summaries of the outcome, such as the median, rather than the mean, and time to disease remission [Phillips et al., 2020]. Alternatively,

Statistica Sinica: Newly accepted Paper

they may summarize several different outcomes rather than just a single one. For example, when learning a treatment allocation, symptom reduction may be considered alongside prognosis [Freemantle et al., 2003]. Most existing approaches for incorporating multiple outcomes involve combining them into a composite outcome and then using policy learning methods designed for single-outcome settings [Butler et al., 2018]. In settings where the actions recommended by experts are recorded in the dataset, Murray et al. provided a means to construct a composite outcome in an automated fashion [Murray et al., 2016]. However, when expert recommendations are not available, composite-outcome-based approaches require investigators to construct the composite outcome in some other way, which often ends up being somewhat arbitrary [Luckett et al., 2021]. Some alternative approaches do not require the construction of a composite outcome. One such approach involves learning a policy that returns a set of recommended actions, rather than a single one [Laber et al., 2014]. Each of the actions in this set should yield a desirable result for at least some of the outcomes.

In cases where a single outcome is of primary interest and others are only of secondary interest, a preferred approach may be to optimize only for this one outcome, while still making inferences about the effect of the policy on the subsidiary outcomes. For example, if a company optimizes a

Statistica Sinica: Newly accepted Paper

policy for customer acquisition, it must also consider the impact the policy will have on customer retention [Afeche et al., 2017]. As another example, just as the side effects of any new medical intervention must be assessed along with its effect on the primary outcome of interest [FDA, 2006], the side effects of a new treatment policy should be assessed as well [Linn et al., 2015]. This problem has been studied in precision medicine where methods have been proposed to learn optimal dynamic treatment regimens with risk constraints [Wang et al., 2018, Liu et al., 2024]. Instead of putting hard constraints on the subsidiary outcomes, this work provides a systematic approach for assessing the impact of a policy that is optimized for some primary outcome, on other subsidiary outcomes.

1.2 Notation and objectives

Let $X \in \mathcal{X}$ be a feature, $A \in \{0, 1\}$ a binary action, and $Y \in \mathcal{Y}$ an outcome that is observed after the action. This outcome may be multivariate. Let $\mathcal{M}$ be a nonparametric model consisting of possible joint distributions P of (X, A, Y). We focus on the offline policy learning setting where the sample consists of n independent and identically distributed draws $(X_i, A_i, Y_i)_{i=1}^n$ from $P_0 \in \mathcal{M}$. Let Π be a set of policies $\mathcal{X} \to \{0, 1\}$ that take as input a feature and take action 0 or 1. For a given policy $\pi \in \Pi$, let $\Omega_\pi(P)$ be a real-

Statistica Sinica: Newly accepted Paper

valued primary performance metric for the policy π under sampling from P, where we assume that larger values of this metric are considered preferable. Further let $\Psi_\pi(P)$ be a real-valued subsidiary performance metric for π under P. For example, when Y is a primary-subsidiary outcome pair $(Y^\star, Y^\dagger) \in \mathbb{R}^2$, these metrics could be the covariate-adjusted means of these two outcomes [Luedtke and Chambaz, 2020, Luedtke and Van Der Laan, 2016]. That is, $\Omega_\pi(P) = \int \mathbb{E}_P[Y^\star|A = \pi(x), X = x]dP(x)$, and $\Psi_\pi(P) = \int \mathbb{E}_P[Y^\dagger|A = \pi(x), X = x]dP(x)$. In this case, let $q : \{0,1\} \times \mathcal{X} \to \mathbb{R}$ such that $q_P(a,x) = \mathbb{E}_P[Y^*|A = a, X = x]$ denotes the value function for the primary metric; and $s : \{0,1\} \times \mathcal{X} \to \mathbb{R}$ such that $s_P(a,x) = \mathbb{E}_P[Y^\dagger|A = a, X = x]$ denotes the value function for the subsidiary metric. Alternatively, the outcome Y may be univariate and the primary performance metric may be equal to the mean $\Omega_\pi(P) = \int \mathbb{E}_P[Y|A = \pi(x), X = x]dP(x)$, while the subsidiary metric may be equal to the covariate-adjusted probability that the outcome exceeds a specified value t, namely $\Psi_\pi(P) = \int P\{Y > t|A = \pi(x), X = x\}dP(x)$. We refer to $\Omega_\pi(P_0)$ and $\Psi_\pi(P_0)$ as the Ω-performance and Ψ-performance of the policy π. For $P \in \mathcal{P}$, let Π^*_P denote the set of optimal policy with respect to the primary performance metric, that is, $\Pi^*_P := \{\pi \in \Pi : \Omega_\pi(P) = \sup_{\pi' \in \Pi} \Omega_{\pi'}(P)\}$. We denote a generic element of this set by π^*_P. We refer to elements of $\Pi^* := \Pi^*_{P_0}$ as Ω-optimal

Statistica Sinica: Newly accepted Paper

policies and denote a generic element by π^*. We assume throughout that Π^* is nonempty, and in general, this set may contain more than one policy.

We are interested in making inferences about the subsidiary performance metric $\Psi_\pi(P_0)$ for Ω-optimal policies. Letting $\psi_0^\ell = \inf_{\pi\in\Pi^*} \Psi_\pi(P_0)$ and $\psi_0^u = \sup_{\pi\in\Pi^*} \Psi_\pi(P_0)$, our objective is to construct a confidence interval for the range of possible Ψ-performances under an Ω-optimal policy, that is, develop a confidence interval that is a superset of $[\psi_0^\ell, \psi_0^u]$ with a specified asymptotic probability.

When there is only one Ω-optimal policy, our objective is to determine the Ψ-performance of this policy, denoted by $\psi_0 := \psi_0^\ell = \psi_0^u$. When there are multiple Ω-optimal policies, ψ_0^ℓ may be less than ψ_0^u, and the upper and lower bounds of our interval inform on the most extreme Ψ-performances that can be attained from an Ω-optimal policy. For example, if larger values of Ψ are preferable, then the upper confidence bound on ψ_0^u informs about the best achievable Ψ-performance by an Ω-optimal policy. Such a policy can be shown to be one of several policies that fall along the Pareto front of the two-objective optimization problem that seeks to maximize Ω and Ψ. The Pareto front denotes the set of policies for which there is not a policy that performs better with respect to one of the two metrics and no worse with respect to the other. The difference between inferring about

ψ_0^u and multi-objective optimization is that the policy with the best Ψ performance is primarily optimized with respect to one performance metric Ω, while multi-objective optimization optimizes several performance metrics simultaneously [Gunantara, 2018, Deb, 2014].

Our main contributions—the first available confidence intervals for subsidiary performance metrics—are presented in the next two sections. When presenting these, we consider two separate cases. In Section 2, we begin with a more specialized case, where the performance metrics Ω_π and Ψ_π are assumed to be the covariate-adjusted means of a primary outcome ($Y^\star$) and subsidiary outcome ($Y^\dagger$). We also assume a unique Ω-optimal policy π^* over an unrestricted policy class Π, and that a margin condition holds. In Section 3, we move to a more general case, where Ω_π and Ψ_π are arbitrary smooth parameters and there may be multiple Ω-optimal policies.

2. Wald-type inference under a margin assumption

In this section, we focus on the case where $\Omega_\pi(P) = \int \mathbb{E}_P[Y^\star|A = \pi(x), X = x]dP(x)$, and $\Psi_\pi(P) = \int \mathbb{E}_P[Y^\dagger|A = \pi(x), X = x]dP(x)$, for a primary-subsidiary outcome pair $(Y^\star, Y^\dagger)$. Moreover, the policy class Π is unrestricted. We aim to build on existing works that evaluate the Ω-performance of an Ω-optimal policy [van der Laan and Luedtke, 2015, Luedtke and Van Der Laan, 2016]. These works have shown that a simple estimation strat-

egy is efficient under a non-exceptional law condition that makes the Ω-optimal rule unique [Robins, 2004]. In this case, $\psi_0^\ell = \psi_0^u$ and we write $\psi_0 = \psi_0^\ell = \psi_0^u$. This strategy first obtains an estimate $\widehat{\pi}$ of the Ω-optimal rule, and then constructs a standard one-step estimator of $\Omega_{\widehat{\pi}}(P_0)$. Heuristically speaking, pursuing estimation of $\Omega_{\widehat{\pi}}(P_0)$, rather than $\Omega_{\pi^*}(P_0)$, introduces only negligible bias because $\widehat{\pi}$ should be a near-maximizer of $\Omega_\pi(P_0)$. Hence, similarly to the fact that $f(x) - f(x^*) = O(|x^* - x|^2)$ for a differentiable function $f : \mathbb{R} \to \mathbb{R}$ with maximizer x^*, the error induced by replacing π^* by $\widehat{\pi}$ in the functional $\pi \mapsto \Omega_\pi(P_0)$ should be of the second-order. In this section, we study the extent to which a standard one-step estimator of $\Psi_{\widehat{\pi}}(P_0)$ will yield an asymptotically normal and efficient estimator of $\Psi(P_0)$. This study is important since, if the standard one-step estimator satisfies these properties under only mild conditions, then there is little reason to develop alternative methods.

We now discuss a key condition that we will require to establish the efficiency of a standard one-step estimator for $\Psi(P_0)$, along with the validity of corresponding Wald-type confidence intervals. Define the function $q_b(P)(x) := \mathbb{E}_P[Y^\star|A = 1, X = x] - \mathbb{E}_P[Y^\star|A = 0, X = x]$ to be the conditional average treatment effect on the primary outcome, and $s_b(P)(x) := \mathbb{E}_P[Y^\dagger|A = 1, X = x] - \mathbb{E}_P[Y^\dagger|A = 0, X = x]$ to be the conditional average

Statistica Sinica: Newly accepted Paper

treatment effect on the subsidiary outcome. We refer to these functions as the primary CATE and subsidiary CATE, respectively. We use the shorthand notation $q_{b,0} := q_b(P_0)$ and $s_{b,0} := s_b(P_0)$.

Condition 1 (Margin condition between $Y^\dagger$ and $Y^\star$)**.** For some $C_1 > 0$ and $\zeta > 2$,

$$P_0\left(|s_{b,0}(X)| \geq C_1 t\,|q_{b,0}(X)|\right) \leq t^{-\zeta}, \qquad \text{for all } t > 1. \tag{2.1}$$

When this condition holds, $|q_{b,0}(X)| \neq 0$ with P_0-probability one. Hence, this condition is a strengthening of the usual non-exceptional law condition [Robins, 2004] that is required when the Ψ and Ω performance metrics coincide. To ensure the validity of the standard one-step estimator, some form of strengthening appears to be needed to make up for the fact that π^* is defined as a maximizer in π of $\Omega_\pi(P_0)$, rather than $\Psi_\pi(P_0)$. Indeed, the estimation error of this estimator $\widehat{\psi}_{\widehat{\pi}}$ can be decomposed as

$$\widehat{\psi}_{\widehat{\pi}} - \Psi_{\pi^*}(P_0) = \left[\widehat{\psi}_{\widehat{\pi}} - \Psi_{\widehat{\pi}}(P_0)\right] + \left[\Psi_{\widehat{\pi}}(P_0) - \Psi_{\pi^*}(P_0)\right].$$

The fact that $\widehat{\psi}_{\widehat{\pi}}$ is a one-step estimator of $\Psi_{\widehat{\pi}}(P_0)$ should imply that the first term will be small. However, since π^* is not necessarily an optimizer for Ψ, it is possible that $\Omega_{\widehat{\pi}}(P_0)$ is close to $\Omega_{\pi^*}(P_0)$ while $\Psi_{\widehat{\pi}}(P_0)$ is far from $\Psi_{\pi^*}(P_0)$ — see Figure 1 for an illustration of this possibility. Therefore, we need a condition to characterize the flatness of the Ψ performance surface

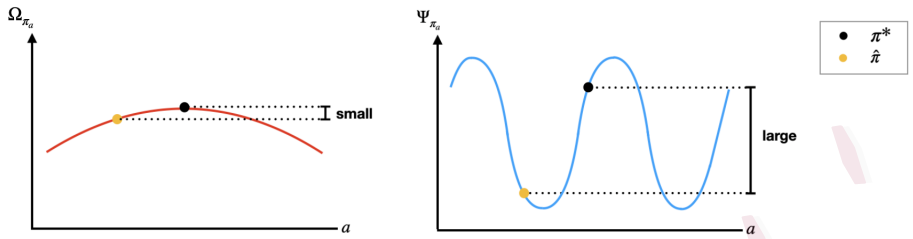

Figure 1: Plot of primary and subsidiary performance metrics for an estimated policy given the threshold policy class $\Pi = \{\mathbf{1}(x \geq a) : a \in \mathbb{R}\}$. The estimator $\widehat{\pi}$ performs well in the sense that the Ω-regret $\Omega_{\pi^*}(P_0) - \Omega_{\widehat{\pi}}(P_0)$ is small, which is to be expected since π^* is defined to be an Ω-optimal rule. Nevertheless, in principle the Ψ-regret $\Psi_{\pi^*}(P_0) - \Psi_{\widehat{\pi}}(P_0)$ could still be large, since the Ψ-value function $\pi \mapsto \Psi_\pi(P_0)$ may be markedly different from the Ω-value function. Though a similar phenomenon can occur for unrestricted policy classes, which are our focus in this section, the infinite-dimensional nature of these classes precludes their visualization.

relative to that of Ω. This flatness can be characterized by studying the absolute CATE ratio $|q_{b,0}(X)| \,/\, |s_{b,0}(X)|$, where we use the convention that $b/0 = +\infty$ for $b > 0$ and we recall that $|q_{b,0}(X)| = 0$ with probability zero under (2.1). Condition 1 imposes that the absolute CATE ratio can only concentrate vanishingly little mass near zero when $X \sim P_0$. This certainly holds in the extreme case where, within each level x of the covariates, the

magnitude of the expected effect of the action on the primary outcome, namely $|q_{b,0}(x)|$, is at least as large as the magnitude of its effect on the subsidiary outcome, namely $|s_{b,0}(x)|$. It also allows for scenarios where the magnitude $|s_{b,0}(x)|$ is much larger than $|q_{b,0}(x)|$ for certain features x with a sufficiently small probability of occurrence. However, it can fail to hold when there are some feature levels where the action does not affect the primary outcome and yet affects the subsidiary outcomes. This can occur, for example, if the primary outcome is cancer remission and the subsidiary outcome captures side effects induced by chemotherapy. Though Condition 1 may be strong, we were unable to show the validity of the standard one-step estimator without it. Therefore, in the remainder of this section, we assume that this condition holds, and we refer the reader to the next section for a method that is valid even when it does not.

In the special case where $Y^\dagger = Y^\star$ a.s., the asymptotic normality and efficiency of the one-step estimator have previously been justified by establishing the pathwise differentiability of the Ω-performance of an Ω-optimal policy [van der Laan and Luedtke, 2015]. We follow a similar approach here when considering cases where $Y^\dagger$ and $Y^\star$ may differ. In particular, we establish the pathwise differentiability of $\Psi^* : P \mapsto \sup_{\pi \in \Pi^*_P} \Psi_\pi(P)$ in what follows. When doing this, we will need to impose Condition 1, along with

an additional margin condition that is inspired by ones previously assumed in policy learning [Qian and Murphy, 2011, Luedtke and Van Der Laan, 2016] and classification [Audibert and Tsybakov, 2007] literature.

Condition 2 (Margin condition for $Y^\star$)**.** For some $\gamma > \frac{1}{\zeta}$,

$$P_0\left(0 < |q_{b,0}(X)| \leq t\right) \lesssim t^\gamma \qquad \forall t > 0. \tag{2.2}$$

This condition imposes that the unique Ω-optimal policy can be estimated well via a plug-in estimator [Qian and Murphy, 2011, Luedtke and Van Der Laan, 2016]. For some generic $P \in \mathcal{P}$ and $\pi \in \Pi$, define $p_P(a|x) := P(A = a|X = x)$ and $D(\pi, P)(x, a, y^\dagger) = \frac{\mathbb{I}\{a=\pi(x)\}}{p_P(a|x)}[y^\dagger - s_P(a, x)] + s_P(\pi(x), x) - \Psi_\pi(P)$. We use the shorthand $p_0 := p_{P_0}$ and $p_n := p_{\widehat{P}_n}$. The following result characterizes the pathwise differentiability of $\Psi^*(\cdot)$ at P_0.

Lemma 1. *Suppose that Ψ_π and Ω_π are covariate-adjusted means for each $\pi \in \Pi$, the policy class Π is unrestricted, and conditions 1 and 2 are satisfied. Then, Ψ^* is pathwise differentiable at P_0 relative to a nonparametric model with canonical gradient $D(\pi^*, P_0)$.*

We use the above result to argue that a one-step corrected estimator is efficient provided its influence function is equal to $D(\Pi^*, P_0)$. Consider

some estimate $\widehat{P}_n$ of the true distribution P_0. The one-step corrected estimator takes the form $\psi_{OS,n} := \Psi_{\widehat{\pi}}(\widehat{P}_n) + P_n D(\widehat{\pi}, \widehat{P}_n)$. For simplicity, when studying this estimator, we focus on the case where $\widehat{\pi}$ is a plug-in estimator of the Ω-optimal policy, namely $\pi^*_{\widehat{P}_n}$. In principle, the policy estimator could be constructed using some other approach, such as outcome-weighted learning [Zhao et al., 2012]. Let $q_{b,n}(x)$ and $s_{b,n}(x)$ be some estimates for the conditional average treatment effects $q_{b,0}(x)$ and $s_{b,0}(x)$ respectively. Also, let $s_n(a, x)$ be some estimate for $s_0(a, x)$. Define the $L_r(P)$ norm of a generic function $f : \mathcal{D} \to \mathbb{R}$ as $\|f\|_{r,P} := [\int_{\mathcal{D}} |f(t)|^r dP(t)]^{1/r}$. We first present some consistency conditions on these estimates.

Condition 3 (Consistent estimator of conditional average treatment effect on the primary outcome)**.** $\|q_{b,n} - q_{b,0}\|^{1+\gamma/2}_{\infty,P_0} = o_{P_0}(n^{-1/2})$.

Condition 4 (Consistent estimator of conditional average treatment effect on the subsidiary outcome)**.** We have

$$\max_{a\in\{0,1\}} \left\{ \left\| \frac{p_0(a \mid \cdot)}{p_n(a \mid \cdot)} - 1 \right\|_{2,P_0} \left\| s_{\widehat{P}_n}(a, \cdot) - s_{P_0}(a, \cdot) \right\|_{2,P_0} \right\} = o_{P_0}(n^{-1/2}).$$

Condition 5 (Donsker function class)**.** $D(\widehat{\pi}, \widehat{P}_n)$ falls in a fixed P_0-Donsker class with probability tending to 1.

Condition 6. $\|D(\widehat{\pi}, \widehat{P}_n) - D(\pi^*, P_0)\|_{2,P_0} \xrightarrow{p} 0$.

Condition 3 guarantees consistency of the conditional average treatment effect estimator and the rate of convergence is slower than the parametric rate when $\gamma > 0$, so this holds as long as we have smoothness and sparsity of the CATE function [Nie and Wager, 2021, Kennedy, 2023]. Condition 4 is standard in the semiparametric inference literature. It is often referred to as an $n^{-1/4}$ rate condition, since it holds if each of two nuisances—in this case $p_0(a \mid \cdot)$ and s_{P_0}—are estimated at that rate [Van Der Laan and Rubin, 2006, Chernozhukov et al., 2018, Kennedy, 2024]; this necessarily holds in regular parametric models, but also holds under enough smoothness in larger models. Condition 5 imposes that the nuisance estimators not be too flexible. This holds, for example, if the inverse propensity and subsidiary outcome regression function estimators fall in function classes of uniformly bounded Hardy-Krause variation with probability one [Benkeser and Van Der Laan, 2016, Fang et al., 2021], since the permanence properties of Donsker classes ensure that this condition is satisfied [Van Der Vaart and Wellner, 2013, Theorem 2.10.6]. Condition 6 requires convergence of the estimated influence functions used for debiasing. Usually, it needs the non-exceptional law condition to hold [Robins, 2004, Luedtke and Van Der Laan, 2016]. The following theorem states that the one-step estimator is efficient.

Theorem 1. *Under Conditions 1–6, the one-step estimator* $\psi_{OS,n}$ *for* $\widehat{\pi} = \pi^*_{\widehat{P}_n}$ *is an asymptotically linear estimator of* $\Psi^*(P_0)$ *with influence function* $D(\pi^*, P_0)$*, in the sense that*

$$\psi_{OS,n} - \Psi^*(P_0) = \frac{1}{n}\sum_{i=1}^{n} D(\pi^*, P_0)(X_i, A_i, Y_i^\dagger) + o_{P_0}(n^{-1/2}).$$

Moreover, $\psi_{OS,n}$ *is an asymptotically efficient estimator of* ψ_0*.*

The above can be used to construct Wald-type confidence intervals for ψ_0 of the form $\psi_{OS,n} \pm z_{1-\alpha/2}\sigma_n/\sqrt{n}$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal random variable and $\sigma_n^2 := \frac{1}{n}\sum_{i=1}^{n} D(\widehat{\pi}, \widehat{P}_n)(X_i, A_i, Y_i^\dagger)^2$.

The Donsker condition required by Theorem 1 can be removed if cross-fitting is used [Schick, 1986]. A 2-fold version of this approach first partitions the data in two halves. Then, it uses the first half of the data to learn $\widehat{\pi}$ and uses the remaining data to construct an estimator for $\Psi_{\widehat{\pi}}(P_0)$. The roles of the halves are then swapped and the two estimators are subsequently averaged. Multi-fold versions of cross-fitting could also be used.

3. Inference of a general functional without margin assumption

3.1 Overview of the methods

The methods we present in this section are agnostic to whether Condition 1 holds and, more generally, whether there are multiple Ω-optimal policies.

Statistica Sinica: Newly accepted Paper

Because the parameter Ψ^* considered in the previous section may not even be well-defined when there are multiple such policies, we instead focus on inferring about the range $[\psi_0^l, \psi_0^u]$ of possible Ψ-performances of Ω-optimal policies. Unlike those in the previous section, the methods developed here critically rely on the policy class Π being restricted — in particular, being P_0-Donsker [Van Der Vaart and Wellner, 2013] — and this condition cannot be removed even if cross-fitting is employed (see Section 3.2 for a discussion). Also, in this section, we do not assume our performance criteria are covariate-adjusted means. Rather, they could take some other form, such as that of a covariate-adjusted median. In what follows we give an overview of our approach for inferring about $[\psi_0^l, \psi_0^u]$.

Our proposed method consists of two stages. The first spends $\beta < \alpha$ error probability to construct a confidence set $\widehat{\Pi}_\beta$ that contains the set of optimal policies Π^* with probability tending to at least $1-\beta$. The second infers about the Ψ-performance of each remaining policy in this confidence set, returning a confidence interval for $[\psi_0^\ell, \psi_0^u]$ of the form

$$\left[\inf_{\pi\in\widehat{\Pi}_\beta}\left\{\widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right\}, \sup_{\pi\in\widehat{\Pi}_\beta}\left\{\widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi z_{\alpha,\beta}}{n^{1/2}}\right\}\right], \tag{3.3}$$

where $\widehat{\psi}_\pi$ is some estimate for $\Psi_\pi(P_0)$, $z_{\alpha,\beta}$ corresponds to the $1-(\alpha-\beta)/2$ quantile of the normal distribution, and $\widehat{\kappa}_\pi^2$ is an estimate of the asymptotic efficiency bound for estimating $\Psi_\pi(P_0)$. We provide a union bounding

Statistica Sinica: Newly accepted Paper

argument that shows that, under conditions, this confidence interval will cover $[\psi_0^\ell, \psi_0^u]$ with asymptotic probability $1-\alpha$.

The first-stage confidence set $\widehat{\Pi}_\beta$ is constructed so that policies that perform poorly in terms of the primary performance metric are eliminated. In other words, we maintain policies π whose uniform upper confidence bound for $\Psi_\pi(P_0)$ is greater than the largest non-uniform lower confidence bound across all policies in the set. Figure 2 shows an example of how the first-stage elimination is performed. More specifically, we define this set $\widehat{\Pi}_\beta$ after the first-stage filtration as

$$\widehat{\Pi}_\beta := \left\{ \pi \in \Pi : L_n \leq \widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} \right\}, \tag{3.4}$$

where $\widehat{\omega}_\pi$ is some estimate for $\Omega_\pi(P_0)$, $\widehat{\sigma}_\pi^2$ is an estimator of the asymptotic efficiency bound for estimating $\Omega_\pi(P_0)$, L_n is an asymptotically valid $1-\beta/2$ lower bound for $\sup_{\pi\in\Pi}\Omega_\pi(P_0)$ (e.g., obtained via [Luedtke and Van Der Laan, 2016]), and t_β is selected in such a way that $\{\widehat{\omega}_\pi + \widehat{\sigma}_\pi t_\beta/n^{1/2} : \pi \in \Pi\}$ is an asymptotically valid $1-\beta/2$ uniform upper confidence bound for $\{\Omega_\pi(P_0) : \pi \in \Pi\}$, in the sense that $\Omega_\pi(P_0) \leq \widehat{\omega}_\pi + \widehat{\sigma}_\pi t_\beta/n^{1/2}$ for all $\pi \in \Pi$ with probability tending to at least $1-\beta/2$ as n goes to infinity. Note that L_n exists and a simple lower bound for L_n is $\sup_{\pi\in\Pi}\left[\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}}\right]$.

It may at first be surprising that, in constructing the confidence interval for $[\psi_0^\ell, \psi_0^u]$, the only place a uniform confidence bound is used is in

Statistica Sinica: Newly accepted Paper

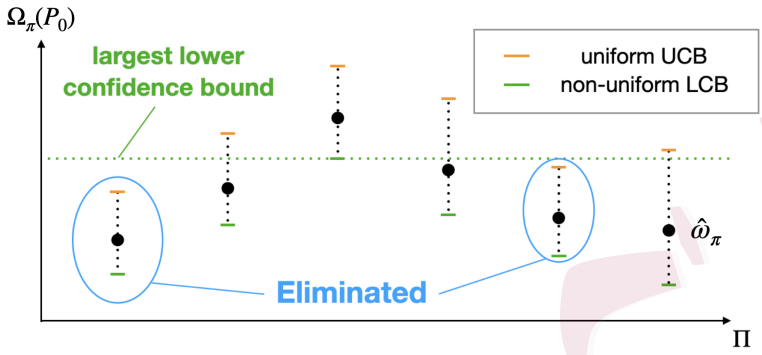

Figure 2: Example of first-stage elimination. Each black dot represents an estimate of $\Omega_\pi(P_0)$ and the horizontal bars denote the confidence bounds. Policies whose uniform upper confidence bound (UCB) is below the largest lower confidence bound (LCB) get eliminated.

the upper bound of (3.4). Indeed, when we began studying this problem, the first approach that we considered was the same as that previously described, except with all confidence bounds replaced by uniform ones. In particular, L_n was defined as the maximum over $\pi \in \Pi$ of a uniform lower confidence bound for the Ω-value function and the minimal and maximal marginal confidence bounds in (3.3) were also replaced by minimal and maximal uniform confidence bounds. However, after analyzing this method, we discovered that less uniformity was needed than we initially expected. Indeed, the uniformity in defining L_n can be dropped since a simple union bounding argument shows that L_n only needs to satisfy that it falls below

Statistica Sinica: Newly accepted Paper

the optimal Ω-value with asymptotic probability at least $1 - \beta/2$; while selecting the maximum of a uniform confidence band for the value function does satisfy such a property, developing such a lower bound is now a well-studied problem, and so less conservative approaches have been developed [Luedtke and Van Der Laan, 2016]. The uniformity in (3.3) can be dropped via an intersection-union method argument [Theorem 1 of Berger and Hsu, 1996], which we show can be applied since our interest concerns parameters defined as the maxima and minima over a set.

As mentioned earlier, justifying the above approach relies on a union-bounding argument across the β coverage error that is made by the first-stage confidence interval in (3.4) and the $1 - \alpha - \beta$ coverage error that is made by the second-stage confidence interval in (3.3). Relying on this union bound could result in unnecessarily wide confidence intervals, so we present another two-stage method whose justification does not require a union bound. In the first stage, we choose the quantiles $s^\dagger_\alpha$, $t^\dagger_\alpha$, and $u^\dagger_\alpha$ derived as extreme values of the joint distributions of estimators of $(\Omega_\pi(P_0))_{\pi\in\Pi}$ and $(\Psi_\pi(P_0))_{\pi\in\Pi}$ — see Section 3.3 for details. Then we construct $\widehat{\Pi}_\beta$ and the asymptotic interval the same ways as in (3.4) and (3.3), while replacing t_β and $z_{\alpha,\beta}$ with $t^\dagger_\alpha$ and $s^\dagger_\alpha$, respectively. Given that $s^\dagger_\alpha$, $t^\dagger_\alpha$, and $u^\dagger_\alpha$ are constructed based on a joint distribution, we refer to this approach as the joint

Statistica Sinica: Newly accepted Paper

approach. Because of the avoidance of the union bound, the joint approach is expected to provide tighter confidence intervals in scenarios when the primary and subsidiary outcomes are strongly correlated.

3.2 A union bounding approach

In this subsection, we provide additional details and theoretical results about the union bounding approach. We first need the following condition for an estimator of $\{\Omega_\pi(P_0) : \pi \in \Pi\}$. In what follows, we let $\tilde{D}_\pi$ be the canonical gradient of Ψ_π relative to a locally nonparametric model, $\sigma_\pi(P_0) := [PD_\pi(P_0)^2]^{1/2}$, and $\kappa_\pi(P_0) := [P\tilde{D}_\pi(P_0)^2]^{1/2}$.

Condition 7 (Uniform asymptotic linearity of estimators of Ω-value and Ψ-value functions)**.** The estimators $\{\widehat{\omega}_\pi : \pi \in \Pi\}$ of $\{\Omega_\pi(P_0) : \pi \in \Pi\}$ and $\{\widehat{\psi}_\pi : \pi \in \Pi\}$ of $\{\Psi_\pi(P_0) : \pi \in \Pi\}$ satisfy

$$\sup_{\pi \in \Pi} [\widehat{\omega}_\pi - \Omega_\pi(P_0) - P_n D_\pi(P_0)] = o_p(n^{-1/2}), \tag{3.5}$$

$$\sup_{\pi \in \Pi^*} \left[\widehat{\psi}_\pi - \Psi_\pi(P_0) - P_n \tilde{D}_\pi(P_0)\right] = o_p(n^{-1/2}). \tag{3.6}$$

These asymptotic linearity conditions can be established via consistency requirements similar to those in Condition 4 and a Donsker condition (see Section 2.1 of [Luedtke and Chambaz, 2020]). Note that (3.6) only requires uniformity over Π^*, rather than all of Π. Estimators satisfying (3.5) and

Statistica Sinica: Newly accepted Paper

(3.6) can be derived via one-step estimation [Pfanzagl, 1982], targeted minimum loss-based estimation [Van Der Laan and Rubin, 2006], or double machine learning [Chernozhukov et al., 2018]. We now provide some conditions on the Ψ-value function, the policy class Π, and necessary conditions for standard deviations and the primary outcome.

Condition 8 (Restricted policy class)**.** The policy class Π satisfies the following:

(1) Π has a bounded uniform entropy integral (Chapter 2.5.1 of [Van Der Vaart and Wellner, 2013]), i.e. $\int_0^\infty \sup_Q \sqrt{\log N(\varepsilon, \Pi, L^2(Q))} d\varepsilon < \infty$, where the sup is over all finitely supported measures Q on $\mathcal{X}$;

(2) Π is closed in $L^2(P_0)$, in the sense that, for all $\pi : \mathcal{X} \to \{0, 1\}$, a Π-valued sequence $(\pi_k)_{k=1}^\infty$ converges to π in $L^2(P_0)$ only if $\pi \in \Pi$;

(3) Π^* is non-empty.

Examples of such policy class Π in $L^2(P_0)$ include classes of binary decision trees with fixed depths while noting that Condition 8 applies to more complicated and general policy classes. We then provide some conditions for the standard deviations and the smoothness of the Ψ-value function.

Condition 9 (Non-vanishing standard deviations and consistent estimators thereof)**.** The following conditions are satisfied: $\inf_{\pi \in \Pi} \sigma_\pi(P_0) > 0$,

Statistica Sinica: Newly accepted Paper

$\sup_{\pi\in\Pi}\sigma_\pi(P_0) < \infty$, $\inf_{\pi\in\Pi}\kappa_\pi(P_0) > 0$, $\sup_{\pi\in\Pi}\kappa_\pi(P_0) < \infty$. In addition, $\widehat{\sigma}_\pi$ and $\widehat{\kappa}_\pi$ are uniformly consistent estimators of $\sigma_\pi(P_0)$ and $\kappa_\pi(P_0)$.

Condition 10 (Smoothness of performance metric in policy)**.** The map $\pi \mapsto \Psi_\pi(P_0)$ is continuous and, for all $\pi, \pi' \in \Pi$, $\|D_\pi - D_{\pi'}\|_{L^2(P_0)} \leq C_2 \|\pi - \pi'\|_{L^2(P_0)}$ for some constant C_2.

When Ω and Ψ are covariate-adjusted mean functionals as in Section 2 and the primary and subsidiary outcomes are bounded, Condition 10 is necessarily true. Let $\mathcal{F} := \{D_\pi(P_0)/\sigma_\pi(P_0) : \pi \in \Pi\}$ and $\tilde{\mathcal{F}} := \{\tilde{f}_\pi := \tilde{D}_\pi(P_0)/\kappa_\pi(P_0) : \pi \in \Pi\}$ denote the collections of canonical gradients that are standardized to have unit variance. Conditions 9 and 10 play a crucial role in showing that $\mathcal{F}$ and $\tilde{\mathcal{F}}$ are P_0-Donsker, which is required to validate the uniform confidence bands utilized in our union bounding approach.

We now show that the confidence set $\widehat{\Pi}_\beta$ defined in (3.4) contains the set of Ω-optimal policies Π^* with high probability asymptotically. Let $\{\mathbb{G}f : f \in \mathcal{F}\}$ be a mean-zero Gaussian process with a covariance function $(f_1, f_2) \mapsto Pf_1f_2$. Then, t_β in (3.4) is defined to be the $1-\beta/2$ quantile of $\sup_{f\in\mathcal{F}}\mathbb{G}f$ and Lemma 4 in the appendix shows that $\left\{\widehat{\omega}_\pi \pm \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}} : \pi \in \Pi\right\}$ is an asymptotically valid uniform β-level confidence band for $\{\omega_\pi : \pi \in \Pi\}$.

Lemma 2 (Asymptotic coverage of $\widehat{\Pi}_\beta$)**.** *If Conditions 7, 8, and 9 hold, then* $\limsup_n P\{\Pi^* \not\subseteq \widehat{\Pi}_\beta\} \leq \beta$.

Statistica Sinica: Newly accepted Paper

The interval in (3.3) uses the remaining $\alpha - \beta$ level of error probability to construct a confidence interval for the random quantity $\widehat{\mathcal{I}}_\beta := [\inf_{\pi \in \widehat{\Pi}_\beta} \Psi_\pi(P_0), \sup_{\pi \in \widehat{\Pi}_\beta} \Psi_\pi(P_0)]$. On the event that $\Pi^* \subseteq \widehat{\Pi}_\beta$, it is true that $\widehat{\mathcal{I}}_\beta \supseteq [\psi_0^\ell, \psi_0^u]$, and so any interval that covers $\widehat{\mathcal{I}}_\beta$ also covers $[\psi_0^\ell, \psi_0^u]$. A union bound then gives our result, which is summarized in Theorem 2.

Theorem 2 (Asymptotic coverage of CI_n)**.** *Under Conditions 7, 8, 9, and 10, for a fixed $\alpha \in (0,1)$ and any choice of $\beta \in (0, \alpha)$, the confidence interval CI_n as defined in (3.3) satisfies $\liminf_{n\to\infty} \mathbb{P}(\{[\psi_0^\ell, \psi_0^u] \subseteq \mathrm{CI}_n\}) \geq 1 - \alpha$.*

Also, as indicated in (3.3), the width of CI_n is determined by a quantile of a standard normal random variable — for example, when $\alpha = 0.06$ and $\beta = 0.01$, $z_{\alpha,\beta} \approx 1.96$. At first, this may seem surprising, given that developing a uniform confidence band for $\{\Psi_\pi : \pi \in \Pi^*\}$ would require using a strictly larger scaling of the standard error of $\widehat{\psi}_\pi$. However, our proof of Theorem 2 shows that using this larger scaling is not necessary for the sake of developing a confidence interval for $[\psi_0^\ell, \psi_0^u]$. The key to this argument involves showing that, under Condition 8, there exist π^ℓ and π^u in Π^* that attain the minimum and maximum Ψ-values, respectively. The existence of π^ℓ shows that the event where the lower bound of CI_n fails to cover $\psi_0^\ell := \inf_{\pi \in \Pi^*} \Psi_\pi(P_0)$, intersected with $\Pi^* \subseteq \widehat{\Pi}_\beta$, satisfies

Statistica Sinica: Newly accepted Paper

$$\left\{ \inf_{\pi\in\Pi^*} \Psi_\pi(P_0) < \inf_{\pi\in\widehat{\Pi}_\beta} \left[\widehat{\psi}_\pi - \widehat{\kappa}_\pi z_{\alpha,\beta}/n^{1/2}\right], \Pi^* \subseteq \widehat{\Pi}_\beta \right\}$$
$$\subseteq \left\{ \inf_{\pi\in\Pi^*} \Psi_\pi(P_0) < \inf_{\pi\in\Pi^*} \left[\widehat{\psi}_\pi - \widehat{\kappa}_\pi z_{\alpha,\beta}/n^{1/2}\right] \right\}$$
$$= \left\{ \psi_{\pi^\ell} < \inf_{\pi\in\Pi^*} \left[\widehat{\psi}_\pi - \widehat{\kappa}_\pi z_{\alpha,\beta}/n^{1/2}\right] \right\} \subseteq \left\{ \psi_{\pi^\ell} < \widehat{\psi}_{\pi^\ell} - \widehat{\kappa}_{\pi^\ell} z_{\alpha,\beta}/n^{1/2} \right\}.$$

The event on the right corresponds to the case where a marginal $1-(\alpha-\beta)/2$-level lower Wald-type confidence interval fails to cover ψ_{π^ℓ}, and so occurs with asymptotic probability $(\alpha-\beta)/2$ under reasonable conditions. In our proof, we establish Theorem 2 using a union bounding argument that combines this with a similar guarantee for the upper bound of CI_n and the fact that $\Pi^* \not\subseteq \widehat{\Pi}_\beta$ happens with asymptotic probability at most β.

Under additional conditions, our confidence interval for $[\psi_0^\ell, \psi_0^u]$ not only ensures asymptotically valid coverage but also attains an optimal $n^{-1/2}$ convergence rate. In this part, we restrict the performance metrics to covariate-adjusted means and propose a boundedness condition on the primary and subsidiary CATE functions.

Condition 11 (Boundedness condition)**.** There exists some $C_3 < \infty$ such that for any $x \in \mathcal{X}$, we have $|s_{b,0}(x)| \leq C_3|q_{b,0}(x)|$.

In most ways, Condition 11 is relatively stronger than Condition 1. Indeed, the limit of Condition 1 as $\zeta \to \infty$ corresponds to the condition that the subsidiary CATE is strictly less than a constant multiple of the

Statistica Sinica: Newly accepted Paper

primary CATE. Since Condition 1 allows for any $\zeta > 2$, it puts a much weaker constraint on how the subsidiary outcome behaves for nearly optimal policies. There is one sense, however, in which Condition 11 is weaker than Condition 1: it does not generally imply that the Ω-optimal policy is necessarily unique. This is true because it allows for equality between subsidiary and primary CATEs, and so both could be zero on some set of positive probability. Though the optimal policy need not be unique when Condition 11 holds, it must still be true that all Ω-optimal policies yield the same Ψ-value, and so in the following lemma we shall let $\psi_0 = \psi_0^\ell = \psi_0^u$.

Lemma 3 ($n^{-1/2}$ convergence rate of CI_n under conditions)**.** *Assume that the performance metrics are covariate-adjusted means as in Section 2, the unrestricted Ω-optimal policy over all possible maps from $\mathcal{X}$ to $\{0,1\}$ is in Π, and $L_n = \sup_{\pi\in\Pi}\left[\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi t_\beta}{n^{1/2}}\right]$ in (3.4). Then, under Conditions 7, 8, 9, and 11, with probability at least $1-2\beta$ asymptotically, the width of the confidence interval for ψ_0 is $O_p(n^{-1/2})$.*

3.3 A joint approach

We now formally describe our joint approach. Consider the mean-zero Gaussian process $\{\mathbb{G}f : f \in \mathcal{F} \cup \tilde{\mathcal{F}}\}$ with covariance function $(f_1, f_2) \mapsto Pf_1f_2$. Our joint approach is the same as the two-stage procedure from

Statistica Sinica: Newly accepted Paper

Section 3.2, except that we require a particular choice of L_n and use cutoffs $(s^\dagger_\alpha, t^\dagger_\alpha, u^\dagger_\alpha)$ satisfying

$$\inf_{\pi\in\Pi} \mathbb{P}\left\{ \inf_{f\in\mathcal{F}} \mathbb{G}f \geq -t^\dagger_\alpha, \sup_{f\in\mathcal{F}} \mathbb{G}f \leq s^\dagger_\alpha, \mathbb{G}\tilde{f}_\pi \geq -u^\dagger_\alpha, \mathbb{G}\tilde{f}_\pi \leq u^\dagger_\alpha \right\} \geq 1-\alpha. \tag{3.7}$$

More specifically, we define the set $\widehat{\Pi}^\dagger$ after the first-stage filtration as

$$\widehat{\Pi}^\dagger := \left\{ \pi \in \Pi : \sup_{\pi\in\Pi} \left[\widehat{\omega}_\pi - \frac{\widehat{\sigma}_\pi s^\dagger_\alpha}{n^{1/2}} \right] \leq \widehat{\omega}_\pi + \frac{\widehat{\sigma}_\pi t^\dagger_\alpha}{n^{1/2}} \right\}. \tag{3.8}$$

Here we choose L_n to be the uppermost point of a uniform lower confidence band for $\{\Omega_\pi(P_0) : \pi \in \Pi\}$ with level $\beta^\dagger := \mathbb{P}\{\sup_{\pi\in\Pi} \mathbb{G}f_\pi > s^\dagger_{1-\alpha}\} < \alpha$. Note that in the union bounding approach, $\beta^\dagger = \beta$, while here $\beta^\dagger$ is implicitly defined through the joint cutoff (3.7). The resulting confidence interval is stated in Theorem 3.

Theorem 3. *Under Conditions 7, 8, 9, 10, assuming the cutoffs* $(s^\dagger_\alpha, t^\dagger_\alpha, u^\dagger_\alpha)$ *satisfy* (3.7)*, it holds that* $\liminf_{n\to\infty} \mathbb{P}(\{[\psi^\ell_0, \psi^u_0] \subseteq \mathrm{CI}^\dagger_n\}) \geq 1-\alpha$*, where*

$$\mathrm{CI}^\dagger_n := \left[\inf_{\pi\in\widehat{\Pi}^\dagger} \left\{ \widehat{\psi}_\pi - \frac{\widehat{\kappa}_\pi u^\dagger_\alpha}{n^{1/2}} \right\}, \sup_{\pi\in\widehat{\Pi}^\dagger} \left\{ \widehat{\psi}_\pi + \frac{\widehat{\kappa}_\pi u^\dagger_\alpha}{n^{1/2}} \right\} \right].$$

There are many possible choices of $(s^\dagger_\alpha, t^\dagger_\alpha, u^\dagger_\alpha)$ that satisfy (3.7). To select among these, we could choose the triple $(s^\dagger_\alpha, t^\dagger_\alpha, u^\dagger_\alpha)$ that provides the tightest confidence interval from this collection, resulting in what we refer to as an optimized joint method. This optimized $(s^\dagger_\alpha, t^\dagger_\alpha, u^\dagger_\alpha)$ is justified since,

Statistica Sinica: Newly accepted Paper

for any choice of $(s_\alpha^\dagger, t_\alpha^\dagger, u_\alpha^\dagger)$ satisfying (3.7), the confidence interval $\mathrm{CI}_{\mathrm{n}}^\dagger$ has valid coverage. For any β, this optimized joint method yields a provably tighter confidence interval than the union bounding method that uses the same choice of L_n as in the left-hand side of (3.8). However, it is possible that the joint approach could potentially result in a wider confidence band in the first stage with the use of an alternative lower confidence bound for the Ω-optimal value, such as the one introduced in Luedtke and Van Der Laan, 2016. In practice, the optimized choice of $(s_\alpha^\dagger, t_\alpha^\dagger, u_\alpha^\dagger)$ is unknown, but it can be approximated via a multiplier bootstrap — see Appendix D for details. Though our theorem focuses on a fixed and known triple $(s_\alpha^\dagger, t_\alpha^\dagger, u_\alpha^\dagger)$, adapting it to allow for the use of an estimated triple with an in-probability limit would be straightforward.

The cutoff in (3.7) considers the joint event regarding $\mathbb{G}\tilde{f}$ and $\mathbb{G}f$ for $f \in \mathcal{F}$ and $\tilde{f} \in \tilde{\mathcal{F}}$, thereby avoiding the use of the union bound required by the approach in Section 3.2. The tightness of this union bound relies on whether the event that Π^* is contained in the first stage policy set, namely $\{\Pi^* \subseteq \widehat{\Pi}_\beta\}$, and the event that $[\psi_0^\ell, \psi_0^u]$ is contained in the second stage confidence interval are disjoint. Of course, when these events are fully disjoint, the union bound will be tight. When they are independent, the (asymptotic) probability that both events occur is $\beta(\alpha - \beta)$, which will be small

for choices of α and β commonly used in practice. Hence, the union bound will only be slightly loose in these cases. Finally, when the events fully overlap, the union bound will be as loose as possible. These scenarios can be better understood by relating them to primary and subsidiary outcomes. Generally, the dependence or independence between the events is likely to correlate with the extent to which primary and subsidiary outcomes depend on each other. The events tend to be independent when primary and subsidiary outcomes are independent, and dependent otherwise.

4. Numerical experiment

We show the performance of our methods on a 1D simulation instance described below. Additional results on a 1D instance with larger sample size, a 3D instance, and two high-dimensional linear instances are in Appendix C.

4.1 A 1D simulation

We conduct simulation studies to evaluate the length and coverage of $1-\alpha$ confidence intervals for bounds on a mean subsidiary outcome, $[\psi_0^\ell, \psi_0^u]$. Our first set of simulations focuses on a 1-dimensional threshold policy class, denoted as $\Pi = \{\mathbf{1}_{[a,\infty)} : a \in [-1, 1]\}$. We compare the confidence intervals from four approaches. The first and the second are the union bounding and

Statistica Sinica: Newly accepted Paper

the joint approaches described in Section 3.2 and 3.3, denoted as union and joint respectively. The third is the one-step estimator approach described in Section 2, denoted as one-step. To ensure that this approach applies, we design our scenarios so that the optimal policy for the unrestricted policy class lies in the threshold class Π. Consequently, in our simulation study, an estimate of the optimal policy in Π also estimates the optimal policy in the unrestricted class. The fourth is a one-step estimator with sample splitting, denoted as os-split. This approach is the same as one-step, except that we obtain an estimate $\widehat{\pi}_1$ of the Ω-optimal policy using only half of the data, and construct a Wald-type confidence interval for $\Psi_{\widehat{\pi}_1}(P_0)$ using the other half. Last, we present an oracle method, denoted as oracle, that knows the specific Ω-optimal policies that provide the upper and lower bounds, ψ_0^u and ψ_0^ℓ. The oracle method uses precisely those policies and construct a Wald-type confidence interval for $[\psi_0^\ell, \psi_0^u]$. Since we have no hope of getting optimal policies *a priori*, the oracle method cannot be used in practice.

We examine three distinct scenarios with an illustration of the Ω and Ψ values of each policy under various scenarios in the three panels in Figure 3. The left panel describes the situation where the set of Ω-optimal policies, Π^*, is not unique. In this scenario, $\Pi^* = \{\mathbf{1}_{[a,\infty)} : a \in [-0.5, 0]\}$, and the margin condition (Condition 1) is not satisfied for any ζ. The middle panel

Statistica Sinica: Newly accepted Paper

describes the situation where Π^* is unique the margin condition is satisfied for any $\zeta > 2$, as we see that when π is around the optimal policy, $q_{b,0}(X)$ varies much faster than $s_{b,0}(X)$. The right panel describes the situation where Π^* is unique but the margin condition is not satisfied for any ζ, as we can see that as X varies, both $q_{b,0}(X)$ and $s_{b,0}(X)$ vary linearly.

For each scenario, we consider sample sizes n of 500 and 5000. To generate the set of policies, we construct a fine grid $(a_1, \cdots, a_N)$ for $N = 10^5$ over $[-1, 1]$ and denote the set of policy as $\Pi_N = \{\mathbf{1}_{[a_i,\infty)} : i \in [N]\}$. We use 1000 multiplier bootstrap replicates to estimate the supremum and infimum in generating the cutoffs. We let $\alpha = 0.05$ when constructing confidence intervals and use 1000 Monte Carlo replications to compute their coverage of the true interval $[\psi_0^\ell, \psi_0^u]$ and approximate their average widths. We estimate the conditional probability $p(a|x)$ via a kernel density estimator as implemented in the sklearn package and the conditional probabilities $p(y|1, x)$ and $p(y|0, x)$ using gradient boosted trees as implemented in the xgboost package, both with the default settings. The Python code to reproduce the simulations is available at https://github.com/zhaoqil/EstimationSubsidiary.

Table 1 shows the coverages and the widths of confidence intervals of $[\psi_0^\ell, \psi_0^u]$ for different scenarios and methods. We can see the one-step estimator fails to provide a nominal coverage when the margin condition

Statistica Sinica: Newly accepted Paper

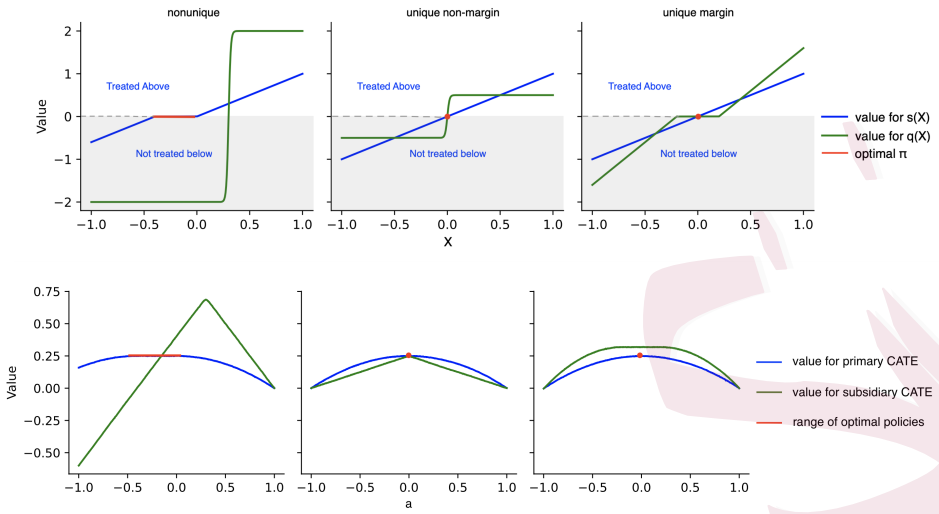

Figure 3: The top figure represents Ω- and Ψ-value, while the bottom figure represents $s_{b,0}(X)$- and $q_{b,0}(X)$-value. 1D threshold policy class $\Pi = \{\mathbf{1}_{[a,\infty)} : a \in [-1, 1]\}$ under different scenarios: 1) the optimal policy for the primary outcome is non-unique, 2) the optimal policy for the primary outcome is unique while the primary and subsidiary outcomes are not correlated, 3) the optimal policy for the primary outcome is unique while the primary and subsidiary outcomes are correlated.

(Condition 1) fails. The other two methods produce similar coverages. We compare the confidence intervals with an oracle confidence interval, which is a lower bound on the width of any valid $1 - \alpha$ confidence interval, and calculate the relative widths. We can see that the joint and union bounding

	coverage				width				
	union	joint	one-step	os-split	union	joint	one-step	os-split	oracle
non-unique	1.000	1.000	0.000	0.000	1.549	1.538	0.240	0.317	0.668
unique non-margin	0.980	0.980	0.812	0.751	0.148	0.143	0.068	0.089	0.068
unique margin	0.978	0.981	0.949	0.953	0.149	0.144	0.074	0.108	0.074

Table 1: Coverages and widths of $[\psi_0^\ell, \psi_0^u]$ with sample size $n = 500$

methods generate confidence intervals about 2.3 times and 2.1 times as wide as the oracle confidence interval when the optimal policy is non-unique and unique, respectively. These results show that although our methods are conservative, they are relatively successful in maintaining a narrow confidence interval. In contrast, the one-step estimator produces a confidence interval that is about the same width as the oracle confidence interval, but it fails to provide valid coverage when the margin condition fails.

5. Discussion

The problem studied in existing works aiming to infer the optimal value of an optimal rule can be viewed as a special case of our setup, where the subsidiary and primary outcomes coincide. In these cases, our two-stage approaches provide ways to make inferences without the margin condition considered in such works [Qian and Murphy, 2011, Luedtke and Van Der Laan, 2016]. Instead, we need uniform asymptotic linearity for the value functions and an appropriately restricted policy class. The margin condition could

fail if the subsidiary metric varies too much across the set of policies that are nearly optimal for the primary metric [Luedtke and Chambaz, 2020]. However, if the policy class is Donsker and the estimator is established via debiased machine learning, the uniform asymptotic linearity condition will be plausible even when a margin condition does not hold.

In our numerical experiments, our union bounding and joint approaches produced valid confidence intervals, even if they were somewhat conservative. Under margin conditions, these intervals attain a parametric $n^{-1/2}$ rate, matching those based on an efficient one-step estimator, although with a less favorable leading constant. However, when the margin conditions fail, intervals based on the one-step estimator fail to achieve valid coverage. In future research, it would be interesting to develop an adaptive procedure that is leading-constant-optimal under margin conditions and, even without them, can produce intervals that provide valid coverage.

Another interesting future direction is to extend our framework to more general constrained policy learning settings, such as risk-constrained or budget-constrained dynamic treatment regimens (DTRs). In these applications, one often wishes to optimize a primary performance measure (e.g., treatment efficacy) while simultaneously satisfying one or more subsidiary constraints (e.g., safety profiles or resource limitations). Our approaches

Statistica Sinica: Newly accepted Paper

could be extended to evaluate or make inferences about these risks.

As for other future work, it is worth exploring methods for inferring subsidiary metrics using observations from adaptive experiments which have a martingale structure. Observations from longitudinal settings could also be considered. Additionally, one could examine simultaneous inference for multiple subsidiary metrics rather than one.

Acknowledgements

This work was supported by National Institutes of Health award DP2-LM013340, and National Science Foundation award DMS-2210216.

Supplementary Materials

Contains proofs of main theorems and lemmas, and additional experiments.

References

P. Afeche, M. Araghi, and O. Baron. Customer acquisition, retention, and service access quality: Optimal advertising, capacity level, and capacity allocation. *Manuf. Serv. Oper. Manag.*, 19(4):674–691, 2017.

S. Athey and S. Wager. Policy learning with observational data. *Econometrica*, 89(1):133–161, 2021.

Statistica Sinica: Newly accepted Paper

REFERENCES

J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Stat.*, 35 (2):608–633, 2007.

D. Benkeser and M. Van Der Laan. The highly adaptive lasso estimator. In *International conference on data science and advanced analytics (DSAA)*, pages 689–696, 2016.

R. L. Berger and J. C. Hsu. Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–319, 1996.

I. Bica, A. M. Alaa, C. Lambert, and M. Van Der Schaar. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clin. Pharmacol. Ther.*, 109(1):87–100, 2021.

S. Boominathan, M. Oberst, H. Zhou, S. Kanjilal, and D. Sontag. Treatment policy learning in multiobjective settings with fully observed outcomes. In *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. & Data Min.*, pages 1937–1947, 2020.

E. L. Butler, E. B. Laber, S. M. Davis, and M. R. Kosorok. Incorporating patient preferences into estimation of optimal individualized treatment rules. *Biometrics*, 74(1):18–26, 2018.

B. Chakraborty, E. B. Laber, and Y. Zhao. Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics*, 69(3):714–723, 2013.

A. Chambaz, W. Zheng, and M. J. van der Laan. Targeted sequential design for targeted learning inference of the optimal treatment rule and its mean reward. *Ann. Stat.*, 45(6), 2017.

Statistica Sinica: Newly accepted Paper

REFERENCES

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1—C68, 2018.

K. Deb. Multi-objective optimization. In *Search methodologies*, pages 403–449. Springer, 2014.

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *The International Conference on Machine Learning (ICML)*, pages 1097–1104, 2011.

B. Fang, A. Guntuboyina, and B. Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy–Krause variation. *The Annals of Statistics*, 49(2):769–792, 2021.

FDA. Guidance for industry: Adverse reactions section of labeling for human prescription drug and biological products – content and format, 2006.

T. Fiez, H. Nassif, Y.-C. Chen, S. Gamez, and L. Jain. Best of three worlds: Adaptive experimentation for digital marketing in practice. In *The Web Conference (WWW)*, pages 3586–3597, 2024.

N. Freemantle, M. Calvert, J. Wood, J. Eastaugh, and C. Griffin. Composite outcomes in randomized trials: greater precision but with greater uncertainty? *JAMA*, 19(289):2554–2559, 2003.

N. Gunantara. A review of multi-objective optimization: Methods and its applications. *Cogent Engineering*, 5(1):1502242, 2018.

Statistica Sinica: Newly accepted Paper

REFERENCES

E. H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *Handbook of Statistical Methods for Precision Medicine*, pages 207–236, 2024.

E. B. Laber, D. J. Lizotte, and B. Ferguson. Set-valued dynamic treatment regimes for competing outcomes. *Biometrics*, 70(1):53–61, 2014.

Y. Ling, P. Upadhyaya, L. Chen, X. Jiang, and Y. Kim. Emulate randomized clinical trials using heterogeneous treatment effect estimation for personalized treatments: Methodology review and benchmark. *Journal of Biomedical Informatics*, 137:104256, 2023.

K. A. Linn, E. B. Laber, and L. A. Stefanski. *Estimation of dynamic treatment regimes for complex outcomes: balancing benefits and risks*, chapter 15, pages 249–262. SIAM, 2015.

L. Liu, Z. Shahn, J. M. Robins, and A. Rotnitzky. Efficient estimation of optimal regimes under a no direct effect assumption. *J. Am. Stat. Assoc.*, 116(533):224–239, 2021.

M. Liu, Y. Wang, H. Fu, and D. Zeng. Learning optimal dynamic treatment regimens subject to stagewise risk controls. *Journal of Machine Learning Research*, 25(128):1–64, 2024.

D. J. Luckett, E. B. Laber, S. Kim, and M. R. Kosorok. Estimation and optimization of composite outcomes. *Journal of Machine Learning Research*, 22(167):1–40, 2021.

A. Luedtke and A. Chambaz. Performance guarantees for policy learning. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 56(3):2162–2188, 2020.

Statistica Sinica: Newly accepted Paper

REFERENCES

A. R. Luedtke and M. J. Van Der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Stat.*, 44(2):713 – 742, 2016.

S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

T. A. Murray, P. F. Thall, and Y. Yuan. Utility-based designs for randomized comparative trials with categorical outcomes. *Statistics in medicine*, 35(24):4285–4305, 2016.

X. Nie and S. Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

J. Pfanzagl. *Contributions to a general asymptotic statistical theory*, volume 13 of *Lecture notes in statistics*. Springer, 1982.

R. Phillips, O. Sauzet, and V. Cornelius. Statistical methods for the analysis of adverse event data in randomised controlled trials: a scoping review and taxonomy. *BMC medical research methodology*, 20(1):1–13, 2020.

M. Qian and S. A. Murphy. Performance guarantees for individualized treatment rules. *Ann. Stat.*, 39(2):1180, 2011.

J. M. Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer, 2004.

A. Schick. On asymptotically efficient estimation in semiparametric models. *Ann. Stat.*, 14(3): 1139–1151, 1986.

Statistica Sinica: Newly accepted Paper

C. Shi, S. Zhang, W. Lu, and R. Song. Statistical inference of the value function for reinforcement learning in infinite horizon settings. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):765–793, 2021.

M. J. van der Laan and A. R. Luedtke. Targeted learning of the mean outcome under an optimal dynamic treatment rule. *Journal of causal inference*, 3(1):61–95, 2015.

M. J. Van Der Laan and D. Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

A. W. Van Der Vaart and J. A. Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer, 2013.

Y. Wang, H. Fu, and D. Zeng. Learning optimal personalized treatment rules in consideration of benefit and risk: with an application to treating type 2 diabetes patients with insulin therapies. *Journal of the American Statistical Association*, 113(521):1–13, 2018.

J. Weltz, T. Fiez, A. Volfovsky, E. Laber, B. Mason, H. Nassif, and L. Jain. Experimental designs for heteroskedastic variance. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 65967–66005, 2023.

B. Zhang, A. A. Tsiatis, E. B. Laber, and M. Davidian. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika*, 100(3):681–694, 2013.

Y. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules

Statistica Sinica: Newly accepted Paper

using outcome weighted learning. *J. Am. Stat. Assoc.*, 107(499):1106–1118, 2012.

Department of Statistics, University of Washington

E-mail: zli9@stanford.edu

Meta Inc.

E-mail: houssamn@meta.com

Department of Statistics, University of Washington

E-mail: aluedtke@uw.edu

Statistica Sinica