OPTIMAL TREATMENT ASSIGNMENT RULES UNDER CAPACITY CONSTRAINTS

KEITA SUNADA AND KOHEI IZUMI

Abstract. We study treatment assignment under capacity constraints, where a planner aims to maximize social welfare by assigning treatments based on observable covariates. Such constraints are common when treatments are costly or limited in supply, but they complicate the analysis of optimal assignment rules because assignment probabilities must be coordinated across the entire covariate distribution. We develop a new approach that reformulates the planner's problem as an *optimal transport* problem, which makes the constraints analytically tractable. Using a limits of experiments framework, we establish local asymptotic optimality results for two canonical decision rules—the plug-in rule and the Bayesian rule. We show that the former rule can dominate the latter rule, with simulations demonstrating sizable risk reductions. An empirical illustration using school voucher program data from Angrist *et al.* (2006) demonstrates how the two rules differ in practice. Our results provide the first decision-theoretic foundation for optimal treatment assignment under capacity constraints.

JEL Classification: C10, C18, C21, C44, C61

Keywords: Treatment assignment, capacity constraints, optimal transport, statistical decision theory, external validity.

1. Introduction

When a social planner allocates treatments such as school vouchers or subsidies, an important question arises: who should be treated, given available data? The planner typically does not know the true treatment effects but can rely on observed data from experimental or observational studies. A growing literature on statistical treatment assignment problems has examined how to use available data to design treatment assignment rules that maximize social welfare. A central goal of this literature is to derive decision-theoretic optimal treatment rules. Two main approaches have been developed toward this goal and continue to be actively studied: the finite-sample approach (e.g., Manski, 2004; Stoye, 2009) and the local asymptotic approach (e.g., Hirano and Porter, 2009).

Despite extensive research on optimal treatment assignment rules, much less is known about how to design optimal rules when the number of available treatments is limited. In many real-world settings, treatment assignment is subject to capacity constraints—budgets, supplies, or available

Date: October 28, 2025 (First version: June 13, 2025).

We are particularly grateful to Nese Yildiz for her guidance and support. We thank Bin Chen, Kazuma Inagaki, Toru Kitagawa, Laura Liu, Yukun Ma, Eric Mbakop, Tai Otsu, Yuya Sasaki, John Singleton, Takeaki Sunada, and the participants at Midwest Econometrics Study Group (UIUC) and Canadian Econometrics Study Group in Ottawa for their helpful comments and suggestions.

slots are limited, so not everyone who could benefit can be treated. This paper extends Hirano and Porter (2009) to derive decision-theoretic optimal rules in the presence of such constraints. At first glance, one might think that the constrained problem is only a straightforward extension of the setting without constraints on available treatments—simply add a quota. Since Hirano and Porter (2009) show that assigning treatment to individuals with (efficiently) estimated positive effects is locally asymptotically optimal when there are no constraints on available treatments, it is tempting to conclude that ranking individuals by their estimated effects and treating them in descending order until the quota is filled would also be optimal. While this intuition turns out to be correct under suitable smoothness conditions on the planner's utility function—a result we establish formally—it is far from straightforward to prove. The presence of capacity constraints makes the derivation of optimal assignment rules substantially more complex, as explained below. This paper develops a new framework using tools from *optimal transport* to accommodate such constraints and establishes decision-theoretic optimal assignment rules—results that, to our knowledge, have not yet been formally derived in the literature. The framework also naturally extends to settings with non-binary treatments.

We consider a planner who aims to maximize social welfare by designing treatment rules based on observable covariates, subject to a capacity constraint. Let $w(\theta, x, t)$ denote the planner's utility from assigning treatment $T = t$ to an individual with covariate $X = x$, where $\theta \in \mathbb{R}^k$ characterizes the treatment effects. For example, if $w(\theta, x, t)$ is the conditional mean of the potential outcome given (x, t), then θ is a parameter indexing the conditional potential outcome distribution. Suppose that the true parameter θ_0 can be efficiently estimated by an estimator $\hat{\theta}_n$ constructed from experimental or observational data. For simplicity, consider a binary treatment setting where a fixed fraction p of the population is to be treated. When θ_0 is known, the planner's problem is

$$\max_{\delta} \int_{\mathcal{X}} \{w(\theta_0, x, 1)\delta(x) + w(\theta_0, x, 0)(1 - \delta(x))\} \, \mathrm{d}F_X(x), \quad \text{s.t.} \int_{\mathcal{X}} \delta(x) \mathrm{d}F_X(x) = p,$$

where F_X is the distribution of covariates and $\delta(x)$ is the probability of assigning treatment to $X = x$. In reality, θ_0 is not known to the planner. Thus the planner utilizes the available data to design treatment assignment rules.

Under the current setup, however, we cannot simply follow a standard approach to analyzing the asymptotically optimal data-driven rules taken by Hirano and Porter (2009) and others. This approach first derives the optimal rule *for each covariate* x in an asymptotically equivalent (and simpler) problem. It then constructs a sequence of feasible data-driven rules that asymptotically match the optimal rule in the limit. The key is that the limiting problem becomes tractable because the optimal rule in the limit is selected from the set of rules obtained as *asymptotic representations* of data-driven rules, which only depend on random variables following a shifted normal and a uniform distribution (van der Vaart, 1991, Theorem 3.1).

Under capacity constraints, this pointwise approach fails because the planner must coordinate assignment probabilities across the entire covariate distribution F_X. Then, one might define the planner's action space as the set of (measurable) functions

$$\mathcal{F} := \left\{ \delta : \mathcal{X} \to [0,1] \,\middle|\, \int_{\mathcal{X}} \delta(x) \mathrm{d}F_X(x) = p \right\},$$

but directly working with $\mathcal{F}$ is inconvenient. It is difficult to obtain the asymptotic representations of assignment rules or to solve the maximization problem because $\mathcal{F}$ fails to be compact under common norms when the support of X is not finite or countable.[1] To proceed, we therefore need a formulation that circumvents this technical difficulty.

Our first contribution is to achieve this by reformulating the planner's constrained problem as an optimal transport problem. The key observation is that, when the capacity constraint binds exactly, the (unconditional) distribution of treatment assignments must be a Bernoulli distribution with success probability p. The planner's problem can then be framed as transporting the mass of F_X into the Bernoulli distribution F_T, in a way that maximizes the social welfare. This reformulation allows us to treat the action space as a set of *couplings*—joint distributions on $\mathcal{X} \times \{0,1\}$ whose marginals are F_X and F_T— equipped with the Wasserstein distance. This new action space is a compact and convex metric space, which are convenient properties for obtaining the asymptotic representations of assignment rules and for solving the planner's problem. Moreover, this framework naturally extends beyond binary treatments to general discrete or continuous settings. To our knowledge, this is the first study to reformulate the optimal treatment assignment problem into an optimal transport problem.

Our second contribution is to provide decision-theoretic optimality results under capacity constraints based on this reformulation. We analyze two canonical assignment rules: the plug-in rule, replaces θ_0 with its efficient estimator $\hat{\theta}_n$, and the Bayesian rule, which forms the posteior on the parameter space Θ based on the observed data. Our main results are as follows: (i) both rules are average optimal under capacity constraints when the planner's utility function is continuously differentiable, (ii) the plug-in rule is generally suboptimal, while the Bayesian rule remains optimal when the planner's utility function is only directionally differentiable.

Directionally differentiable utility functions often arise in empirically relevant settings. For instance, they arise when the (conditional) potential outcome distributions may differ slightly between the target and training populations, and the planner adopts a maximin-type utility function that is robust to such distributional shifts (Adjaho and Christensen, 2023).

Our results are consistent with the existing literature. The plug-in rule is known to be average optimal under point-identified models when the planner's utility function is continuously differentiable (Hirano and Porter, 2009), while the Bayesian rule is known to be average optimal under

[1]By van der Vaart (1991, Theorem 3.1), $\mathcal{F}$ needs to be complete and separable to obtain the asymptotic representations of assignment rules. Under the sup-norm, $\mathcal{F}$ is typically not separable when $\mathcal{X}$ is not countable. Under L_p-norm; i.e., $\|\cdot\|_p = \left(\int |\cdot|^p \,\mathrm{d}F_X(x)\right)^{1/p}$, $\mathcal{F}$ is separable and complete, but not compact unless $\mathcal{X}$ is finite. Thus it is not straightforward to find a suitable norm to work with $\mathcal{F}$.

partially identified models when the planner's utility function is directionally differentiable (Christensen *et al.*, 2025), where the lack of point identification similarly precludes full differentiability and renders the plug-in rule suboptimal.[2] Our results extend these findings to settings with capacity constraints, providing the first decision-theoretic optimality results under such constraints.

To quantify the performance gap between the two rules, we conduct a simulation study. The results support our theoretical findings. Specifically: (i) the Bayesian rule achieves a substantially lower risk than the plug-in rule in small samples for both continuously and directionally differentiable utility functions, (ii) the two rules behave similarly under the continuously differentiable utility function in larger samples, and (iii) the Bayesian rule continues to outperform the plug-in rule under the directionally differentiable utility function in larger samples.

We illustrate our methods using data from Angrist *et al.* (2006), who study the impact of receiving a randomly assigned voucher—allowing students to attend private high schools—on educational attainment seven years later. We hypothetically treat the marginal distribution of covariates (age and sex) in the observed sample as that of the target population, and compute both the plug-in rule and the Bayesian rule. The two rules yield identical allocations under a continuously differentiable utility function but differ under the maximin-type directionally differentiable utility function mentioned above.

1.1. **Related literature.** This study builds on the literature on the statistical treatment assignment problems in econometrics, where the pioneering works include Manski (2004) and Dehejia (2005). Within this expanding literature, our main contribution is to provide a decision-theoretic optimality result under capacity constraints on available treatments. Previous studies have established the decision-theoretic optimality of treatment rules in several settings including: (i) point-identified smooth (semi-)parametric models under local asymptotics (Hirano and Porter, 2009; Masten, 2023), (ii) partially-identified smooth (semi-)parametric models under local asymptotics (Christensen *et al.*, 2025; Kido, 2023; Xu, 2024), (iii) point-identified models with finite samples (Stoye, 2009; Stoye, 2012; Tetenov, 2012; Guggenberger *et al.*, 2024; Kitagawa *et al.*, 2024; Chen and Guggenberger, 2025), (iv) partially-identified models with finite samples (Manski, 2007; Stoye, 2012; Yata, 2023; Ishihara and Kitagawa, 2024; Aradillas Fernández *et al.*, 2024; Montiel Olea *et al.*, 2024). However, none of these studies consider the capacity constraints in the way we do. To the best of our knowledge, this is the first study to establish an optimality result under such constraints within the framework of Hirano and Porter (2009), extended to cover non-binary treatments—both discrete and continuous.

Besides Hirano and Porter (2009), the most closely related paper is Christensen *et al.* (2025). They extend Hirano and Porter (2009) to allow for partially identified parameters and (non-randomized) discrete actions, and show the asymptotic optimality of the Bayesian rule. Our setting differs in that the action space is the set of couplings of F_X and F_T (see notation below) equipped

[2]Formally, Christensen *et al.* (2025) adopt an optimality criterion that is weaker than average optimality. Under additional assumptions their results imply average optimality as well.

with the Wasserstein distance, which naturally allows for randomized assignment rules. This nonstandard action space requires nontrivial extensions to analyze the asymptotic properties of the Bayesian rule, which we address by drawing on tools from optimal transport. In contrast to their framework, we focus on point-identified parameters.[3]

There also exist studies that incorporate exogenously given constraints into the treatment assignment problem. Bhattacharya and Dupas (2012) impose the capacity constraints in the way we do, but focus on the estimation and inference of a nonparametric plug-in rule. Other papers adopting the *empirical welfare maximization* approach allow for various types of constraints, including the capacity constraints (Kitagawa and Tetenov, 2018; Athey and Wager, 2021; Mbakop and Tabord-Meehan, 2021; Sun, 2024). Kitagawa and Tetenov (2018) show the optimality of their proposed rule in terms of the welfare convergence rate, which measures how quickly the average welfare achieved by the proposed rule converges to the maximum welfare under the true data generating process.

Some recent works utilize tools from optimal transport theory in the literature of treatment assignment problems. Kido (2022) and Adjaho and Christensen (2023) study the external validity of treatment choices by measuring the difference in potential outcome distributions between the training and target populations using the Wasserstein distance. Hazard and Kitagawa (2025) formulate a learning problem of optimal matching policies in a two-sided market as an empirical optimal transport problem, and derive a welfare regret bound for their estimated policy.

Our work also relates to the growing field of statistical methods for optimal transport problems (Chewi *et al.*, 2024), as we study the local asymptotic properties of transport maps of an optimal transport problem where the cost function is indexed by parameters that can be efficiently estimated.

1.2. **Structure of the paper.** The remainder of the paper is organized as follows. Section 2 formulates the planner's problem and introduces the data generating process. Section 3 introduces the decision theoretic framework and define the plug-in rule and the Bayesian rule. Then the optimality results are stated. Section 4 provides a simulation study to evaluate the finite sample performance of rules. Section 5 illustrates our methods using the data from Angrist *et al.* (2006). Finally, Section 6 concludes. All of the proofs are relegated to Appendix.

1.3. **Notation.** A function $f : \Theta \subset \mathbb{R}^k \to \mathbb{R}$ is *(Hadamard) directionally differentiable* at θ_0 if there is a continuous function $\dot{f}_{\theta_0} : \mathbb{R}^k \to \mathbb{R}$ such that

$$\lim_{n\to\infty} \left| \frac{f(\theta_0 + t_n h_n) - f(\theta_0)}{t_n} - \dot{f}_{\theta_0}(h) \right| = 0$$

for all sequences $\{t_n\} \subset \mathbb{R}_+$ and $\{h_n\} \subset \mathbb{R}^k$ such that $t_n \downarrow 0$, $h_n \to h \in \mathbb{R}^k$ as $n \to \infty$ and $\theta_0 + t_n h_n \in \Theta$ for all n. It is worth noting that this requires $\dot{f}_{\theta_0}$ need to be continuous, but not to be linear.

[3] Xu (2024) further extends the framework of Christensen *et al.* (2025) to continuous decision problems using an expansion-based approach.

Let P and Q be Borel probability measures on $\mathcal{A}$ and $\mathcal{B}$, respectively. A joint distribution μ on $\mathcal{A} \times \mathcal{B}$ is called a *coupling* of P and Q if its marginals are P and Q; that is, $\mu(A \times \mathcal{B}) = P(A)$ and $\mu(\mathcal{A} \times B) = Q(B)$ for any measurable sets A and B.

2. Setting

The setting of this paper closely follows that of Hirano and Porter (2009). We consider a social planner who assigns a treatment T to individuals based on their observable covariates X. Let F_X denote the marginal distribution of X in the target population, with support $\mathcal{X}$. We assume that F_X is known to the planner. The planner can fractionally (probabilistically) assign treatment $T = t$ to an individual with covariate $X = x$.

Let $Y(t)$ denote potential outcomes under treatment $T = t$. In contrast to Hirano and Porter (2009), we distinguish between the conditional potential outcome distribution in the target population and in the training population. We denote the conditional distribution of $Y(t)$ in the training population as $F_t(\cdot|x, \theta)$, where $F_t(\cdot|x, \theta)$ belongs to families of distributions indexed by a parameter $\theta \in \Theta \subset \mathbb{R}^k$. The planner must learn θ from the available data from experimental or observational studies.

2.1. The planner's preferences.

2.1. The planner's preferences. The planner's utility for assigning treatment $T = t$ to an individual with covariate $X = x$ depends on the conditional distribution $F_t(\cdot|x, \theta)$ via a functional w. For the shorthand notation, we write

$$w(\theta, x, t) := w(F_t(\cdot|x, \theta)).$$

We consider two scenarios. First, $w(\theta, x, t)$ is fully differentiable in θ. Second, $w(\theta, x, t)$ is only directionally differentiable in θ. Two examples corresponding to each scenario are given as follows.

Example 2.1. When the planner is interested in the (conditional) mean outcome, then a natural choice of utility function is

$$w(\theta, x, t) = \int y \mathrm{d}F_t(y|x, \theta).$$

This choice is standard in the literature and appropriate especially when the target and training populations are assumed to have the same conditional potential outcome distribution. ■

Example 2.2. When the conditional potential outcome distributions may differ slightly across the target and training populations, the planner may wish to adopt a utility function that is robust to distributional shifts. Following Adjaho and Christensen (2023), we formalize such a utility function.[4] We first define an ε-neighborhood of $F_t(\cdot|x, \theta)$ as

$$\mathcal{N}_\varepsilon := \left\{ G_t(\cdot|x) : \tilde{d}_W(G_t(\cdot|x), F_t(\cdot|x, \theta)) \le \varepsilon \right\},$$

[4]For alternative approaches to robust welfare, see Si *et al.* (2020), Kido (2022), and Qi *et al.* (2023).

where $\varepsilon > 0$ is a measure of neighborhood size, and $\tilde{d}_W$ is the Wasserstein distance of order 1.[5] Given $\lambda \in [0,1]$, the planner's utility for assigning treatment $T = t$ for individuals with covariate $X = x$ is then defined as

$$w(\theta, x, t) = \lambda \int y \mathrm{d}F_t(y|x,\theta) + (1-\lambda) \inf_{G_t(\cdot|x) \in \mathcal{N}_\varepsilon} \int y \mathrm{d}G_t(y|x).$$

This formulation corresponds to maxmin preferences of an ambiguity-averse decision maker (Gilboa and Schmeidler, 1989). In the second term, since the true target distributions are unknown, the planner computes the conditional mean under the worst-case distribution within the neighborhood of $F_t(\cdot|x,\theta)$, treating it as a fixed reference prior on $Y(t)$ given x.[6]

By Adjaho and Christensen (2023, Remark 2.2), this utility function can be rewritten as

$$w(\theta, x, t) = \lambda \int y \mathrm{d}F_t(y|x,\theta) + (1-\lambda) \max\left\{ \int y \mathrm{d}F_t(y|x,\theta) - \varepsilon, y_\ell(t) \right\}, \tag{2.1}$$

where $y_\ell(t)$ is the possible minimum values in the support of $Y(t)$. From this expression, one finds that (2.1) is only directionally differentiable, and the directional derivative of the second term of (2.1) for direction h at θ is given by $(1-\lambda)$ times

$$\begin{cases} \left(\frac{\partial}{\partial \theta} \int y \mathrm{d}F_t(y|x,\theta) \right)^\top h & \text{if } \int y \mathrm{d}F_t(y|x,\theta) - \varepsilon > y_\ell(t), \\ \max\left\{ \left(\frac{\partial}{\partial \theta} \int y \mathrm{d}F_t(y|x,\theta) \right)^\top h, 0 \right\} & \text{if } \int y \mathrm{d}F_t(y|x,\theta) - \varepsilon = y_\ell(t), \\ 0 & \text{if } \int y \mathrm{d}F_t(y|x,\theta) - \varepsilon < y_\ell(t). \end{cases}$$

We note that the non-linearity of this directional derivative corresponds to the failure of the full differentiability of (2.1) at θ. ■

2.2. **The planner's problem as optimal transport.** We consider a setting where the planner faces the capacity constraints on the available treatments. To illustrate, consider a simple binary treatment case, $T \in \mathcal{T} := \{0,1\}$, where a fraction p of the target population is to be treated. We assume that the capacity constraint binds exactly. Let $\delta(x)$ denote the probability of assigning treatment $T = t$ to individuals with $X = x$. Then, under the capacity constraint, the planner's problem can be written as

$$\max_{\delta \in \mathcal{F}} \int_{\mathcal{X}} \{ w(\theta, x, 1)\delta(x) + w(\theta, x, 0)(1 - \delta(x)) \} \mathrm{d}F_X(x), \tag{2.2}$$

We now show how to convert this constrained optimization problem into more tractable one. Observe that the distribution of treatment assignments must be a Bernoulli distribution with the success probability p. With this in mind, the planner's problem can be seen as an optimal transport

[5] Formally,

$$\tilde{d}_W(F_t(\cdot|x), G_t(\cdot|x,\theta)) := \inf_{\pi \in \Pi(F_t, G_t)} \int |y - \tilde{y}| \, \mathrm{d}\pi(y, \tilde{y}),$$

where $\Pi(F_t, G_t)$ denotes all couplings of F_t and G_t. Note that $\tilde{d}_W$ differs from d_W defined in (2.3).

[6] Several recent studies incorporate non-Bayesian preferences that arise naturally in settings with set-identifiable parameters (Giacomini and Kitagawa, 2021; Aradillas Fernández *et al.*, 2024; Christensen *et al.*, 2025). Banerjee *et al.* (2020) adopt maxmin preferences to study of the optimal experimental design problems.

problem: the planner transports the mass of F_X into F_T, the Bernoulli distribution, in a way that maximizes the social welfare.

Formally, let $\mathcal{M}_a$ be the set of all couplings of F_X and F_T. Let d be a distance function on $\mathcal{X} \times \mathcal{T}$, and define the Wasserstein distance of order 1 as

$$d_W(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \int d((x,t),(x',t'))\mathrm{d}\gamma((x,t),(x',t')), \tag{2.3}$$

where $\Gamma(\mu, \nu)$ is the set of couplings whose marginals are μ and ν. We focus on couplings that have a finite first moment:

$$\mathcal{M} := \left\{ \mu \in \mathcal{M}_a : \int d((x_0,t_0),(x,t))\mathrm{d}\mu < +\infty \right\},$$

for some arbitrary $(x_0, t_0) \in \mathcal{X} \times \mathcal{T}$. Then $(\mathcal{M}, d_W)$ becomes a metric space, and d_W is finite on $\mathcal{M}$ (Villani, 2009, Theorem 6.9). Using this setup, the original constrained problem (2.2) is equivalent to the following:

$$\max_{\mu \in \mathcal{M}} W(\theta, \mu), \tag{2.4}$$

where

$$W(\theta, \mu) := \int_{\mathcal{X} \times \mathcal{T}} w(\theta, x, t)\mathrm{d}\mu = \int_{\mathcal{X}} \int_{\mathcal{T}} w(\theta, x, t)\mathrm{d}\mu(t|x)\mathrm{d}F_X(x).$$

Hence, the action space of the planner is the space of couplings $(\mathcal{M}, d_W)$.

There are three important remarks regarding this optimal transport formulation. First, the capacity constraint is automatically satisfied by any coupling in $\mathcal{M}$, making the problem effectively unconstrained. Second, this reformulation is computationally attractive, as one can use an existing software for optimal transport. Finally, this approach can easily accommodate non-binary treatment settings. We assume that T follows a distribution F_T, determined by the capacity constraints, with support $\mathcal{T}$. We remark that $\mathrm{d}\mu(t|x)$ becomes a conditional probability measure when F_T is continuous.

2.3. **The data generating process.** When the true (finite-dimensional) parameter θ_0 is known to the planner, the optimal rules can be obtained by solving

$$\max_{\mu \in \mathcal{M}} W(\theta_0, \mu).$$

However, since the planner does not know θ_0 in practice, she must select a rule in a data-driven manner. For this purpose, data $Z^n = (Z_1, \ldots, Z_n)$, which are informative about θ (and hence about $F_t(\cdot|x, \theta)$) are available from a training population. We assume that the data Z^n are i.i.d. with $Z_i \sim P_\theta$ on some space $\mathcal{Z}$ equipped with the Borel σ-algebra $\mathcal{B}(\mathcal{Z})$. We let P_θ^n denote the joint probability measure of Z^n. In Appendix E, we consider an extension in which the sampling distribution of data may depend on (possibly infinite-dimensional) nuisance parameters, as in a GMM model.

Example 2.3. Angrist *et al.* (2006) study the medium-term effects of the PACES program in Columbia.[7] Specifically, they investigate the impact of receiving a randomly assigned voucher (which allows attendance at private schools) on educational attainment seven years later, using the administrative data.

In one of their main specifications, they consider the following linear model:

$$Y_i = X_i^\top \beta + \alpha T_i + u_i,$$

where Y_i denotes test scores, X_i includes observable covariates (sex and age), T_i is an indicator for the treatment status, and u_i is an error term. However, not all individuals in the sample took the exam. Because the voucher recipients were more likely to take the exam than non-recipients, a selection issue arises. To address this, Angrist *et al.* (2006) construct a modified test score variable by censoring the observed scores at a specific quantile of the test score distribution. Let R_i be an indicator for exam registration and $\tau > 0$ be the censoring threshold. Then the censored test score is defined as

$$Y_i(\tau) := \mathbf{1}\{R_i Y_i \geq \tau\} Y_i + \mathbf{1}\{R_i Y_i < \tau\} \tau.$$

Under the assumptions that (i) u_i is normally distributed with mean zero, and (ii) any untested student would have scored below the threshold τ had they taken the exam, the parameters can be consistently estimated using a Tobit-type maximum likelihood estimator. In this context, the parameter is $\theta = (\alpha, \beta, \sigma)$, and the observed data is $Z^n = \{(T_i, X_i, Y_i(\tau)) : i = 1, \ldots, n\}$, where σ is the standard deviation of the error term u_i. ■

Following Hirano and Porter (2009) and among others, we use a local asymptotic framework where we perturb the data-generating process around the true one. Let Θ be an open subset of $\mathbb{R}^k$ and suppose that θ_0 is the true parameter. Let $\theta_{nh} := \theta_0 + h/\sqrt{n}$. We assume that the sequence of experiments $\mathcal{E}_n = \{P_\theta^n : \theta \in \Theta\}$ satisfies *differentiability in quadratic mean* (DQM) at θ_0: there exists a function $s : \mathcal{Z}^n \to \mathbb{R}^k$ such that

$$\int \left[\mathrm{d}P_{\theta_0+h}^{1/2}(z) - \mathrm{d}P_{\theta_0}^{1/2}(z) - \frac{1}{2} h' s(z) \mathrm{d}P_{\theta_0}^{1/2}(z) \right]^2 = o(\|h\|^2) \quad \text{as } h \to 0, \tag{2.5}$$

where s is the score function associated with $\mathcal{E}_1$. Let $I_0 = \mathbb{E}_{P_{\theta_0}^n}[ss']$.

The planner's *statistical treatment assignment rule* (or just *rule*) $\mu : \mathcal{Z}^n \to \mathcal{M}$ maps realizations of data into the coupling. Let

$$A_0 := \arg\max_{\mu \in \mathcal{M}} W(\theta_0, \mu)$$

be the set of couplings that maximize the welfare at θ_0. We define the class of sequences of rules by

$$\mathcal{D} := \left\{ \{\mu_n\} : \mu_n(Z^n) \overset{h}{\rightsquigarrow} Q_{\theta_0,h} \text{ and } \sqrt{n} P_{\theta_{nh}}^n \left(\mu_n(Z^n) \notin A_0\right) \to 0 \quad \forall h \in \mathbb{R}^k, \forall \theta_0 \in \Theta \right\}, \tag{2.6}$$

where $\overset{h}{\rightsquigarrow}$ denotes convergence in distribution along $P_{\theta_{nh}}^n$ with $Z^n \sim P_{\theta_{nh}}^n$ for each n, and $Q_{\theta_0,h}$ is a (possibly degenerate) probability measure on $\mathcal{M}$. For technical reasons, we restrict our analysis to

[7] PACES stands for Programa de Ampliación de Cobertura de la Educación Secundaria.

rules satisfying $\sqrt{n}P^n_{\theta_{nh}}(\mu_n(Z^n) \notin A_0) \to 0$, a condition also imposed by Christensen *et al.* (2025). This condition ensures that the treatment rule maximizes the welfare at the true parameter θ_0 with high probability in a neighborhood of θ_0, and that the probability of selecting a suboptimal coupling (i.e., $\mu_n \notin A_0$) vanishes sufficiently fast.

Since $(\mathcal{M}, d_W)$ is compact (and thus complete and separable) by Villani (2009, Remark 6.19), we obtain the following asymptotic representation theorem by van der Vaart (1991, Theorem 3.1).

Proposition 2.4. *Let $\{\mu_n\} \in \mathcal{D}$ satisfy $\mu_n \overset{h}{\rightsquigarrow} Q_{\theta_0,h}$ for all $h \in \mathbb{R}^k$ and $\theta_0 \in \Theta$. Under the DQM condition, there exists $\mu_\infty : \mathbb{R}^k \times [0,1] \to \mathcal{M}$ such that for every $h \in \mathbb{R}^k$, $\mathcal{L}_h(\mu_\infty(\Delta, U)) = Q_{\theta_0,h}$, $\mathcal{L}_h(\Delta) = N(h, I_0^{-1})$, and $\mathcal{L}_h(U) = \mathrm{Unif}[0,1]$ with $U \perp\!\!\!\perp \Delta$.*

This proposition states that any sequence $\{\mu_n\}$ in $\mathcal{D}$ can be matched by some treatment rule μ_∞ in a *limit experiment* where we observe a single draw Δ from a shifted normal distribution and an independent uniform random variable U. This representation is useful for analyzing the asymptotic properties of rules, as the limit experiment is more analytically tractable than the original experiments $\mathcal{E}_n$.

Remark 2.5. Hirano and Porter (2009) study the optimality of treatment assignment rules after conditioning on a fixed covariate value $X = x$. In our notation, this is equivalent to finding an optimal conditional probability $\mu(\cdot|x)$ for each x. Accordingly, they derive the asymptotic representation of $\mu(\cdot|x)$ as a function of Z^n by applying a version of the representation theorem (see their Proposition 3.1). However, because it is difficult to accommodate the capacity constraints within this framework, we instead apply the representation theorem to couplings $\mu \in \mathcal{M}$. Note that in our setup, the map $Z^n \mapsto \mu(Z^n)$ takes values in $\mathcal{M}$, rather than the unit interval.

3. Optimal decisions

3.1. **Decision theoretic framework and rules.** We begin by introducing a decision theoretic framework to evaluate the performance of a sequence of rules $\{\mu_n\} \in \mathcal{D}$. Let

$$W^*_{\mathcal{M}}(\theta) := \max_{\mu' \in \mathcal{M}} W(\theta, \mu')$$

denote the maximum attainable welfare at θ. Following the literature, we define the *welfare regret* $W^*_{\mathcal{M}}(\theta) - W(\theta, \mu)$ as the loss incurred from choosing a coupling $\mu \in \mathcal{M}$. Accordingly, the *risk* associated with the map $Z^n \mapsto \mu(Z^n) \in \mathcal{M}$ at θ is given by

$$R(\mu, \theta) := \mathbb{E}_{P^n_\theta}\left[W^*_{\mathcal{M}}(\theta) - W(\theta, \mu(Z^n))\right], \tag{3.1}$$

where the expectation is taken with respect to the sampling distribution P^n_θ of Z^n. The planner's objective is to minimize the risk by constructing data-driven rules $\{\mu_n\} \in \mathcal{D}$.

Let π be any prior density function on Θ that is continuous and positive at θ_0. A sequence of rules $\{\mu^*_n\} \in \mathcal{D}$ is said to be *average optimal* if $\{\mu^*_n\}$ attains the infimum of the asymptotic risk function:

$$\inf_{\{\mu_n\}\in\mathcal{D}} \liminf_{n\to\infty} \int \sqrt{n} R(\mu_n, \theta_{nh})\pi(\theta_{nh})\mathrm{d}h. \tag{3.2}$$

Our main goal is to construct a sequence of rules that is average optimal. A natural candidate is the plug-in rule, which has been shown to be average optimal by Hirano and Porter (2009) in settings without capacity constraints. To formalize this rule, let $\hat{\theta}_n$ be a *best regular estimator* of θ_0 such that

$$\sqrt{n}(\hat{\theta}_n - \theta_{nh}) \overset{h}{\rightsquigarrow} N(0, I_0^{-1}) \quad \text{as } n \to \infty.$$

The maximum likelihood estimator or the Bayesian posterior mean estimator are typical examples of best regular estimators. The *plug-in rule* is the sequence $\{\mu_n^P(Z^n)\}$, where for each n,

$$\mu_n^P(Z^n) \in \arg\max_{\mu\in\mathcal{M}} W(\hat{\theta}_n, \mu).$$

Another rule we examine is the *Bayesian rule* $\{\mu_n^B(Z^n)\}$, defined for each n by

$$\mu_n^B(Z^n) \in \arg\max_{\mu\in\mathcal{M}} \int_\Theta W(\theta, \mu)\pi_n(\theta|Z^n)\mathrm{d}\theta,$$

where $\pi_n(\theta|Z^n)$ is the posterior density obtained from a strictly positive, continuous prior density π on Θ. Christensen *et al.* (2025) show the average optimality of such Bayesian rules in discrete choice problems when (i) the model is partially identified, and (ii) decision rules are not fractional.

3.2. **An optimality result for the Bayesian rule.** We provide an optimality result for a Bayesian rule under the directional differentiability of w. We discuss asymptotic properties of the plug-in rule in Subsection 3.3. We impose the following assumptions for results in this subsection. As in Clarke and Barron (1990) and Christensen *et al.* (2025), we say that a family $\mathcal{P}$ is *locally quadratic* if for any $\theta_0 \in \Theta$, $D_{\mathrm{KL}}(p_\theta \parallel p_{\theta'}) \le 2(\theta-\theta')^\top I_0(\theta-\theta')$ holds for any θ and θ' belonging to a neighborhood of θ_0, where $D_{\mathrm{KL}}(p_\theta \parallel p_{\theta'})$ is the Kullback-Leibler divergence with respect to a common dominating measure ν. Also, we say $\mathcal{P}$ is *sound* if weak convergence of P_θ to $P_{\theta'}$ is equivalent to the convergence of θ to θ' for probability measures $P_\theta, P_{\theta'}$ and parameters $\theta, \theta' \in \Theta$.

Assumption 3.1. *(i) Θ is open. (ii) $\mathcal{P}$ is DQM at any θ_0. (iii) I_0 is finite and nonsingular at any θ_0. (iv) $\mathcal{P}$ is locally quadratic. (v) $\mathcal{P}$ is sound.*

The first three conditions are standard assumptions in local asymptotic frameworks. Conditions (iv) and (v) ensure that Schwartz's theorem—which originally establishes posterior consistency in a space of density functions (see, for example, Ghosh and Ramamoorthi (2003) and Ghosal and van der Vaart (2017))—can be applied in a parametrized setting, as in Clarke and Barron (1990). We impose these assumptions to show that the Bayesian rule μ_n^B satisfies $\sqrt{n}P_{\theta_{nh}}^n\left(\mu_n^B(Z^n) \notin A_0\right) \to 0$. The same conditions are also imposed by Christensen *et al.* (2025).

Assumption 3.2. *$\mathcal{X} \times \mathcal{T}$ is compact in the product metric space where the distance function d is equipped.*

For example, this condition is satisfied if $\mathcal{X}$ is a compact metric space, and $\mathcal{T}$ is a finite discrete space.

Assumption 3.3. *(i) $w(\theta, x, t)$ is bounded continuous on Θ for any $(x, t) \in \mathcal{X} \times \mathcal{T}$. (ii) $w(\theta, x, t)$ is continuous on $\mathcal{X} \times \mathcal{T}$ uniformly over Θ.*

Note that discrete covariates are compatible with condition (ii), since the metric d can be defined to incorporate the discrete metric.

Assumption 3.4. *(i) $w(\theta, \cdot)$ is directionally differentiable (as a function on $\mathcal{X} \times \mathcal{T}$) at any $\theta \in \Theta$ with derivative $\dot{w}_\theta$.*[8] *(ii) $\dot{w}_{\theta_0}(x, t; h)$ is continuous on $\mathcal{X} \times \mathcal{T}$ for any h. (iii) $\dot{w}_{\theta_0}(\cdot; h)$ is uniformly dominated by a function $K(h)$ that grows at most subpolynomially of order p; i.e., $\max_{(x,t)\in\mathcal{X}\times\mathcal{T}} |\dot{w}_{\theta_0}(x, t; h)| \leq K(h) \leq 1 + \|h\|^p$.*

Condition (i) imposes a uniform version of directional differentiability, rather than requiring it only pointwise in (x, t). For condition (iii), a similar polynomial growth condition appears in the study of Bayes estimators (van der Vaart, 1998, Section 10.3). Choosing $p = 1$ is sufficient for (2.1) provided $\max_{(x,t)\in\mathcal{X}\times\mathcal{T}} \left\| \frac{\partial}{\partial\theta} \int y \mathrm{d}F_t(y|x, \theta_0) \right\| < \infty$.

Assumption 3.5. *(i) The prior (Lebesgue) density function π is positive, continuous, and bounded on Θ . (ii) $\int \|\theta\|^p \pi(\theta)\mathrm{d}\theta < \infty$.*

The order p used in condition (ii) must align with the order in Assumption 3.4 (iii).

Assumption 3.6. *There exists K such that for all $(\mu, \nu) \in A_0 \times (\mathcal{M} \setminus A_0)$, $W(\theta_0, \mu) > K \geq W(\theta_0, \nu)$.*

Note that $W(\theta_0, \mu)$ is constant over A_0. This condition requires that the value of $W(\theta_0, \cdot)$ is uniformly separated between A_0 and $\mathcal{M} \setminus A_0$. The requirement arises because $\mathcal{M}$ is infinite; it is unnecessary when the action space is finite. To see an implication from this condition, note that the correspondence $A(\theta) := \arg\max_{\mu\in\mathcal{M}} W(\theta, \mu)$ is upper hemicontinuous at θ_0 by the theorem of maximum of Berge. From this observation, one can show that Assumption 3.6 implies that for sufficiently small $\varepsilon > 0$ we have that $A(\theta) = A_0$ for all $\theta \in N_\varepsilon(\theta_0)$, which means that $A(\theta)$ is invariant around the neighborhood of θ_0.

We note that the sequence $\{\mu_n^B(Z^n)\}$ of Bayesian rules may not be uniquely determined as our framework allows for multiple maximizers of the objective function. This non-uniqueness complicates the analysis since we cannot directly apply the argmax theorem, which is often used to study the asymptotic behavior of general argmax-functionals.

[8]That is, for any $r_n \downarrow 0$ and $h_n \to h$,

$$\max_{(x,t)\in\mathcal{X}\times\mathcal{T}} \left| \frac{w(\theta_0 + r_n h_n, x, t) - w(\theta_0, x, t)}{r_n} - \dot{w}_{\theta_0}(x, t; h) \right| \to 0.$$

To deal with this, we utilize a penalized version of the Bayesian rule. Let $\nu \in \mathcal{M}$ be any fixed reference measure and $H : \mathcal{M} \to \mathbb{R}_+$ be a functional given by $\mu \mapsto (d_W(\mu,\nu))^2$, which will serve as a penalty function of a maximization problem. For example, we can let ν be the product measure of F_X and F_T. The functional H has following properties:

Proposition 3.1. *H is a nonnegative, continuous, strictly convex, and bounded functional on $(\mathcal{M}, d_W)$.*

Proof. See Appendix C. □

Then we define the penalized Bayesian rule by

$$\mu^B_{n,\varepsilon}(z) := \arg\max_{\mu\in\mathcal{M}} \int \sqrt{n} W(\theta,\mu)\pi_n(\theta|z)\mathrm{d}\theta - \varepsilon H(\mu),$$

for $\varepsilon > 0$. Note that $\mu^B_{n,\varepsilon}$ becomes the unique maximizer of this penalized problem by the strict convexity of H. We then obtain a useful result on the penalized rules $\{\mu^B_{n,\varepsilon}\}$ by following the arguments of Nutz (2022).[9]

Proposition 3.2. *Let $\mathcal{M}_{opt}(z) := \arg\max_{\mu\in\mathcal{M}} \int \sqrt{n} W(\theta,\mu)\pi_n(\theta|z)\mathrm{d}\theta$. For each $z \in \mathcal{Z}^n$, there exists a unique $\mu^B_n(z) \in \mathcal{M}$ such that (i) $\mu^B_{n,\varepsilon}(z)$ converges weakly to $\mu^B_n(z)$ as $\varepsilon \downarrow 0$, (ii) $\mu^B_n(z) \in \mathcal{M}_{opt}(z)$, and (iii) $\mu^B_n(z) = \arg\min_{\mu\in\mathcal{M}_{opt}(z)} H(\mu)$.*

Proof. See Appendix D. □

Thus, we can construct a unique $\{\mu^B_n(z)\}$ where $\mu^B_n(z)$ minimizes the penalty function H over $\mathcal{M}_{opt}(z)$. The following result is stated in terms of $\{\mu^B_n(Z^n)\}$ defined in this way.

Theorem 3.3. *Under Assumptions 3.1–3.6, $\{\mu^B_n\} \in \mathcal{D}$ is average optimal.*

Proof. See Appendix A. □

Our proof strategy follows the approaches of Hirano and Porter (2009), Christensen *et al.* (2025), and Xu (2024), but requires suitable extensions since our action space $(\mathcal{M}, d_W)$ is more complicated than theirs. This gives rise to technical challenges specific to our framework, which we address by drawing on tools from optimal transport.

Remark 3.4. Formally, the construction of $\{\mu^B_n\}$ depends on the choice of a prior distribution. However, any prior satisfying Assumption 3.5 leads to the same conclusion.

Remark 3.5. $\{\mu^B_n(Z^n)\}$ remains average optimal if the directional differentiability in Assumption 3.4 is strengthened to the full differentiability.

[9]Nutz (2022) provides corresponding results by choosing H as the Kullback-Leibler (KL) information criterion between $\mu \in \mathcal{M}$ and any reference measure in $\mathcal{M}$. KL is nonnegative and strictly convex in μ, but not continuous and bounded. Here we impose stronger requirements for the penalty function H, which is needed to handle weak convergence of functionals on $\mathcal{M}$ to study the asymptotic properties of rules. Accordingly, the mode of convergence of $\mu^B_{n,\varepsilon}(z)$ is modified to weak convergence from convergence in total variation, see Nutz (2022, Theorem 5.5).

Remark 3.6. Christensen *et al.* (2025) impose a condition called *no first-order ties*, which requires the uniqueness of the minimizer of the loss function in the limit experiment. This condition addresses an indeterminacy: their treatment rule must be matched to one of the minimizers, but it is not uniquely determined without this condition. By contrast, our construction of the rule $\{\mu_n^B(Z^n)\}$ allows us to avoid imposing this condition, since $\mu_n^B(z)$ is matched with μ_∞^* in the limit experiment where μ_∞^* uniquely minimizes the penalty function H over the set of maximizers of (3.3) defined below.[10]

3.3. **Asymptotic behavior of the plug-in rules.** We will explore the asymptotic behavior of the plug-in rules to compare with the Bayesian rule.

3.3.1. *When w is directionally differentiable.* As a part of the proof of Theorem 3.3, we show that any average optimal rule $\{\mu_n(Z^n)\} \in \mathcal{D}$ will be matched by the rule in the limit experiment μ_∞ where μ_∞ satisfies

$$\mu_\infty(\Delta) \in \arg\max_{\mu \in A_0} \int \int \dot{w}_{\theta_0}(x,t;s) \mathrm{d}N(\Delta, I_0^{-1})(s) \mathrm{d}\mu, \tag{3.3}$$

with $\Delta \sim N(h, I_0^{-1})$ (see Lemma A.2). The Bayesian rule satisfies this condition. We want to check whether the plug-in rule $\mu_n^P(Z^n)$ satisfies this.

One can show that $\{\mu_n^P\} \in \mathcal{D}$, which implies $\mu_n^P(Z^n) \in A_0$ with probability approaching to one along $P_{\theta_{nh}}^n$. Thus, for sufficiently large n, $\mu_n^P(Z^n)$ equivalently solves

$$\begin{aligned} \arg\max_{\mu \in A_0} \sqrt{n} W(\hat{\theta}_n, \mu) &= \arg\max_{\mu \in A_0} \int \sqrt{n} w(\hat{\theta}_n, x, t) \mathrm{d}\mu \\ &= \arg\max_{\mu \in A_0} \sqrt{n} \left[\int w(\hat{\theta}_n, x, t) \mathrm{d}\mu - \int w(\theta_0, x, t) \mathrm{d}\mu \right] \end{aligned}$$

where the second equality follows because the value of $W(\theta_0, \mu)$ is constant across $\mu \in A_0$. We will see that the maximization problem that plug-in rules solve weakly converges to a different maximization problem from the one that the matched rules of optimal rules must solve; i.e., (3.3). We actually claim that

$$\max_{\mu \in A_0} \sqrt{n} \left[\int w(\hat{\theta}_n, x, t) \mathrm{d}\mu - \int w(\theta_0, x, t) \mathrm{d}\mu \right] \overset{h}{\rightsquigarrow} \max_{\mu \in A_0} \int \dot{w}_{\theta_0}(x,t;\Delta) \mathrm{d}\mu \quad \text{as } n \to \infty, \tag{3.4}$$

where $\Delta \sim N(h, I_0^{-1})$.

To see this, let $B_n(\mu) := \sqrt{n} \left[\int w(\hat{\theta}_n, x, t) \mathrm{d}\mu - \int w(\theta_0, x, t) \mathrm{d}\mu \right]$ and $B_\infty(\mu) := \int \dot{w}_{\theta_0}(x,t;\Delta) \mathrm{d}\mu$. We impose a high-level condition that the process $\{B_n(\mu) : \mu \in A_0\}$ is asymptotically tight. Also note that $\sqrt{n}(\hat{\theta} - \theta_0) = I_0^{-1} S_n + o_{P_{\theta_0}^n}(1)$ with $S_n \overset{0}{\rightsquigarrow} N(0, I_0)$ as $n \to \infty$ by the best regularity of $\hat{\theta}_n$. Combining Le Cam's third lemma and the delta method for the directionally differentiable functions (Fang and Santos, 2019, Theorem 2.1) yields

$$\sqrt{n} \left[\int w(\hat{\theta}_n, x, t) \mathrm{d}\mu - \int w(\theta_0, x, t) \mathrm{d}\mu \right] \overset{h}{\rightsquigarrow} \int \dot{w}_{\theta_0}(x,t;\Delta) \mathrm{d}\mu \quad \text{as } n \to \infty,$$

[10] Xu (2024) assumes uniqueness of the rule and therefore does not encounter the issue of non-uniqueness we addressed here.

where $\Delta \sim N(h, I_0^{-1})$. By the asymptotic tightness of $\{B_n(\mu) : \mu \in A_0\}$, we can extend this result to convergence in distribution of the process

$$B_n \overset{h}{\rightsquigarrow} B_\infty \quad \text{as } n \to \infty \text{ on } \ell^\infty(A_0),$$

by van der Vaart and Wellner (1996, Theorem 1.5.4). Then applying the continuous mapping theorem yields

$$\max_{\mu \in A_0} B_n(\mu) \overset{h}{\rightsquigarrow} \max_{\mu \in A_0} B_\infty(\mu) \quad \text{as } n \to \infty,$$

which completes the argument.

It is evident that the solutions of RHS of (3.4) need not to solve (3.3), the maximization problem in the limit experiment that the matched rules of optimal rules must solve. Thus the plug-in rules might not be average optimal in general when w is directionally differentiable. We provide a concrete example in which the the matched rule of the plug-in rule does not solve (3.3) in the following example.

Example 3.7. Assume that the planner observes a single covariate X, say sex, with $\mathcal{X} = \{x_f, x_m\}$. Suppose $F_X(x_f) = F_X(x_m) = 1/2$. Consider binary treatments setup where a fraction $p = 1/2$ of the individuals to be treated. Assume that the conditional mean of the potential outcome is given by

$$\int y \mathrm{d}F_t(y|x, \theta) = \begin{cases} t\theta^2 & \text{if } x = x_f, \\ t\theta & \text{if } x = x_m, \end{cases} \qquad \theta_0 = 1.$$

Thus, the mean outcome is identical for x_f and x_m at true value θ_0, but differs locally around θ_0. Further, let $\varepsilon > 0$ satisfy $-\varepsilon < y_\ell(0)$. In this case, the directional derivative of the maximin utility function defined in (2.1) with $\lambda = 0$ is

$$\dot{w}(x_f, 1; h) = \begin{cases} 2h & \text{if } h \geq 0, \\ 0 & \text{if } h < 0, \end{cases} \qquad \dot{w}(x_f, 0; h) = 0,$$

and

$$\dot{w}(x_m, 1; h) = \begin{cases} h & \text{if } h \geq 0, \\ 0 & \text{if } h < 0, \end{cases} \qquad \dot{w}(x_m, 0; h) = 0.$$

Note that $\mathcal{M} = A_0$ since $\theta_0 = 1$.

Under this setup, we have

$$\int\int \dot{w}_{\theta_0}(x, t; s) \mathrm{d}N(\Delta, I_0^{-1})(s) \mathrm{d}\mu = \frac{1}{2} h(\Delta)\mu(1|x_f) + h(\Delta)p, \quad h(\Delta) := \left[\Delta - \frac{\phi(-\Delta)}{1 - \Phi(-\Delta)}\right],$$

where ϕ and Φ denote the pdf and cdf of the standard normal distribution, respectively. Let Δ_h satisfy $h(\Delta_h) = 0$ (numerically, $\Delta_h \approx 0.506$). Then note that $h(\Delta) \geq 0$ if $\Delta \geq \Delta_h$ and $h(\Delta) < 0$ if

$\Delta < \Delta_h$. Therefore, the optimal rule in (3.3) is

$$\begin{cases} \mu(1|x_f) = 1, \mu(1|x_m) = 0 & \text{if } \Delta \geq \Delta_h, \\ \mu(1|x_f) = 0, \mu(1|x_m) = 1 & \text{if } \Delta < \Delta_h. \end{cases}$$

In contrast, we obtain

$$\int \dot{w}_{\theta_0}(x, t; \Delta) \mathrm{d}\mu = \left[\max\{\Delta, 0\} - \max\left\{\frac{\Delta}{2}, 0\right\} \right] \mu(1|x_f) + \max\{\Delta, 0\} p.$$

Therefore, the optimal rule in (3.4) is

$$\begin{cases} \mu(1|x_f) = 1, \mu(1|x_m) = 0 & \text{if } \Delta \geq 0, \\ \mu(1|x_f) = 0, \mu(1|x_m) = 1 & \text{if } \Delta < 0, \end{cases}$$

which is suboptimal for $0 \leq \Delta \leq \Delta_h$.

Remark 3.8. To investigate the asymptotic properties of the plug-in rule formally, we need to handle the non-uniqueness issue. This can be done by the penalization used for the construction of the Bayesian rule.

3.3.2. *When w is fully differentiable.* If we strengthen the directional differentiability in Assumption 3.4 to the full differentiability, then the plug-in rules become average optimal. To see this, notice that $\dot{w}_{\theta_0}(x, t; s) = \dot{w}_{\theta_0}(x, t)^\top s$ for some $\dot{w}_{\theta_0}(x, t) \in \mathbb{R}^k$ from the linearity of the directional derivative. Then the maximization problem (3.3) is rewritten as

$$\max_{\mu \in A_0} \int \dot{w}_{\theta_0}(x, t)^\top \Delta \mathrm{d}\mu,$$

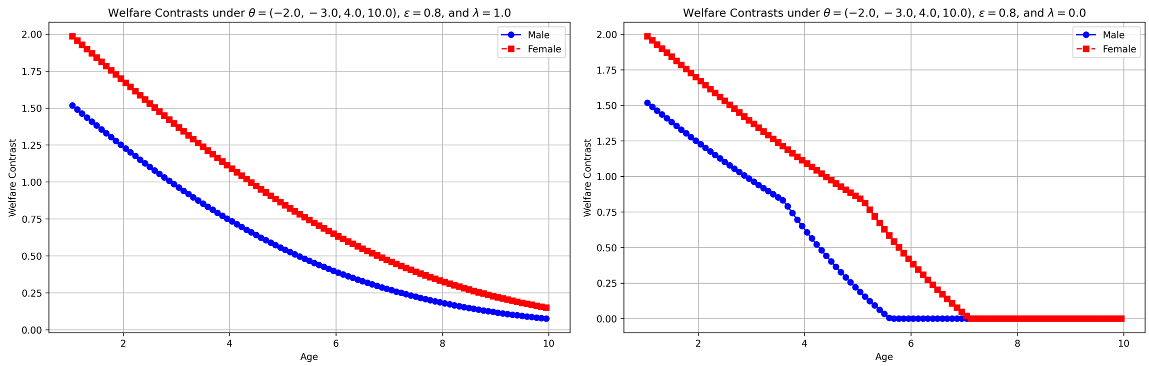
which is the same as (3.4), the maximization problem that the matched rules of the plug-in rules solve in the limit experiment.

This pattern is consistent with findings from the existing literature. As discussed earlier, Hirano and Porter (2009) show that the plug-in rule is average optimal in point-identified models when the utility function is fully differentiable. In contrast, Christensen *et al.* (2025) demonstrate that the Bayesian rule is average optimal in partially identified models when the utility function is only directionally differentiable, and the plug-in rule fails to be optimal unless the full differentiability holds. In partially identified settings, directional differentiability is a natural and often unavoidable assumption, as full differentiability typically does not hold.

4. Simulation

We conduct a simulation study to evaluate the performance of the Bayesian rule and the plug-in rule under the following conditions: (i) the welfare function is either smooth or only directionally differentiable, and (ii) the sample size is relatively small ($n = 200$) and large ($n = 500$).

Figure 1. Welfare contrasts under smooth and directionally differentiable welfare at θ_0



We closely follows the data generating process described in Example 2.3. For the training population, the latent variable is generated by

$$Y_i^* = X_i^\top \beta + \alpha T_i + u_i,$$

where $X_i \in \mathbb{R}^2$ denotes the observable covariates and T_i is the binary treatment that is randomly assigned. The first coordinate of X_i, interpreted as age, follows a truncated normal distribution with mean 4, standard deviation of 2, and is bounded on $[1, 10]$. The second coordinate, interpreted as sex, is a binary variable assigned with equal probability. The observed outcome is

$$Y_i = \max\{0, Y_i^*\}.$$

We set $\beta_0 = (-2, -3)$, $\alpha_0 = 4$, and $u_i \sim N(0, \sigma_0^2)$ with $\sigma_0 = 10$. The observed data is an i.i.d. sample $Z^n = \{(Y_i, X_i, T_i)\}_{i=1}^n$. The parameters $\theta_0 = (\beta_0, \alpha_0, \sigma_0)$ can be estimated by the maximum likelihood using Z^n.

In this Tobit model, the conditional mean of the potential outcomes in the training population is given by

$$w(\theta, x, t) = (x^\top \beta + \alpha t) - (x^\top \beta + \alpha t)\Phi\left(\frac{-x^\top \beta - \alpha t}{\sigma}\right) + \sigma \phi\left(\frac{-x^\top \beta - \alpha t}{\sigma}\right),$$

where Φ and ϕ denote the standard normal cdf and pdf, respectively. Following Examples 2.1 and 2.2, we specify the planner's utility function as

$$w_R(\theta, x, t, \varepsilon, \lambda) = \lambda w(\theta, x, t) + (1 - \lambda) \max\{w(\theta, x, t) - \varepsilon, 0\},$$

where $\lambda \in [0, 1]$ and $\varepsilon > 0$. We note that w_R is differentiable when $\lambda = 1$, but only directionally differentiable otherwise. In what follows, we focus on the cases $\lambda = 0, 1$ and $\varepsilon = 0.8$.

Figure 1 plots the welfare contrasts $w_R(\theta, x, 1, \varepsilon, \lambda) - w_R(\theta, x, 0, \varepsilon, \lambda)$ for both males and females at θ_0. Under $\lambda = 0$, kinks appear when $w(\theta, x, t) - \varepsilon < 0$. For a given covariate x, the contrast is zero if both $w(\theta, x, 1) < \varepsilon$ and $w(\theta, x, 0) < \varepsilon$; that is, individuals with sufficiently low welfare are regarded as deriving no benefit from treatment.

The upper panels of Figure 4 show the oracle (infeasible optimal) rule at θ_0. The rule assigns treatment to females younger than approximately 6.5 and to males younger than approximately 5. Notably, this oracle rule remains unchanged across $\lambda = 0$ and $\lambda = 1$. It also remains the same at $\theta_0 + h/\sqrt{n}$, for the range of local deviation parameters h specified below.

We assume that the true distribution of covariates X in the training population is known and that the target population shares the same distribution. Specifically, we define F_X as the joint distribution of (a) the truncated normal distribution for the age variable and (b) the binary distribution for the sex variable. For computational purposes, we discretize F_X into 99 bins, each corresponding to a distinct combination of age and sex. Each bin is assigned a probability mass according to F_X, representing the proportion of individuals falling into that bin.

Suppose that the planner has resources to allocate to 75% of the target population. Let

$$W(\theta, \mu) = \int w_R(\theta, x, t, \varepsilon, \lambda)\mathrm{d}\mu(x, t).$$

4.1. **Average optimality.** We evaluate performance under a sequence of perturbed DGPs:

$$\theta_{nh} = \theta_0 + h/\sqrt{n}, \quad \text{for } h \in H := \{-2, -1.6, \ldots, 2\},$$

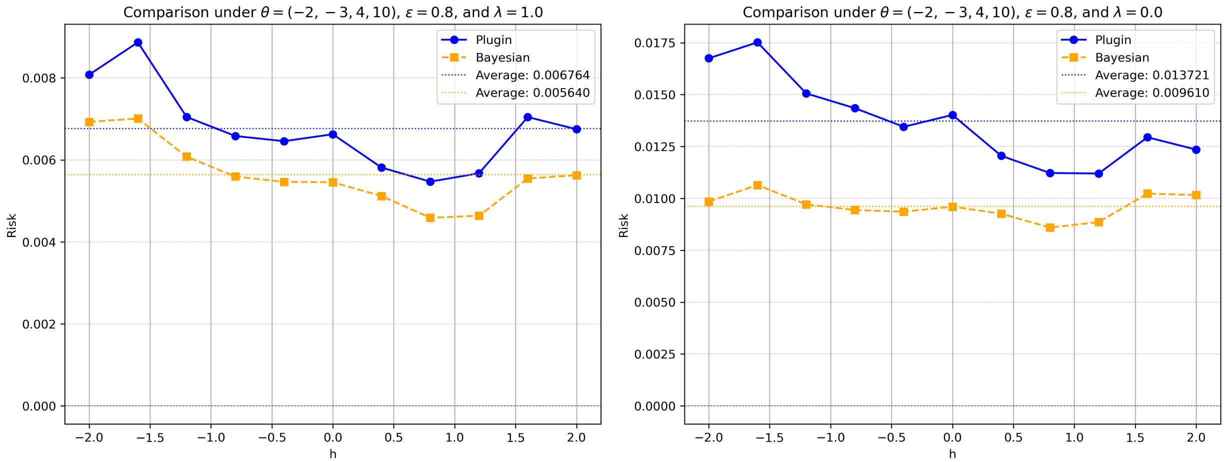
where $h/\sqrt{n}$ is added to θ_0 element-wisely. Our goal is to compare the average risk

$$\int R(\mu_n^Q, \theta_{nh})\mathrm{d}h = \int \mathbb{E}_{P^n_{\theta_{nh}}}\left[W^*_{\mathcal{M}}(\theta_{nh}) - W(\theta_{nh}, \mu_n(Z^n))\right]\mathrm{d}h$$

for $Q = P, B$. The simulation proceeds as follows:

(1) For each h, draw J independent samples of data $\{Z^{n,j}\}_{j=1}^J$ from $P^n_{\theta_{nh}}$ where $Z^{n,j} = \{Z_i^j\}_{i=1}^n$.
(2) For each j:
 (a) Obtain the MLE estimates of parameter $\hat{\theta}^j_{nh}$ and Fisher information matrix $\hat{I}^j_{nh}$.
 (b) Compute the plug-in rule $\mu^{P,j}_{nh}$ by
$$\mu^{P,j}_{nh} \in \arg\max_{\mu \in \mathcal{M}} W(\hat{\theta}^j_{nh}, \mu).$$
 (c) Draw L samples $\{\theta_\ell\}_{\ell=1}^L$ from $N(\hat{\theta}^j_{nh}, (n\hat{I}^j_{nh})^{-1})$, which can be interpreted as the quasi-posterior using the quasi-likelihood $N(\hat{\theta}^j_{nh}, (n\hat{I}^j_{nh})^{-1})$ with the uniform prior (Kim, 2002; Christensen *et al.*, 2025). Then compute the Bayesian rule $\mu^{B,j}_{nh}$ by
$$\mu^{B,j}_{nh} \in \arg\max_{\mu \in \mathcal{M}} \frac{1}{L}\sum_{\ell=1}^{L} W(\theta_\ell, \mu).$$
(3) Compute the oracle welfare $W^*_{\mathcal{M}}(\theta_{nh}) = \max_{\mu \in \mathcal{M}} W(\theta_{nh}, \mu)$, and estimate $R(\mu_n^Q(Z^n), \theta_{nh})$ by
$$R(Q, h) := \frac{1}{J}\sum_{j=1}^{J}\left[W^*_{\mathcal{M}}(\theta_{nh}) - W(\theta_{nh}, \mu^{Q,j}_{nh})\right]$$
for $Q = P, B$. Store $R(Q, h)$ for each h.
(4) Taking the average of $R(Q, h)$ over h gives an estimate of $\int R(\mu_n^Q, \theta_{nh})\mathrm{d}h$ for $Q = P, B$.

FIGURE 2. Comparisons of estimated risks: $n = 200$ (left: smooth welfare, right: directionally differentiable welfare)



We use POT, an open-source Python library developed by Flamary *et al.* (2021), to compute the plug-in rule and the Bayesian rule.

4.2. **Results.** We study the cases of $n = 200, 500$, $J = 2000$, and $L = 2000$ for both $\lambda = 0$ and $\lambda = 1$. We first report the estimated risks, followed by comparisons of the resulting treatment allocations.

Figure 2 shows the results for $n = 200$. While our theory predicts the plug-in and Bayesian rules are asymptotically optimal under smooth welfare ($\lambda = 1$), the simulation shows that the Bayesian rule performs better in small samples. We also observe that the Bayesian rule outperforms the plug-in rule under directionally differentiable welfare ($\lambda = 0$).

Figure 3 shows the results for $n = 500$. The Bayesian rule still performs slightly better when $\lambda = 1$, but the overall risk levels are substantially reduced, and the performance gap between the two rules narrows. This indicates that both rules are approaching optimality as the sample size increases from 200 to 500. When $\lambda = 0$, the Bayesian rule continues to outperform the plug-in rule, which is consistent with our theoretical predictions: under the directionally differentiable welfare the Bayesian rule is optimal, but the plug-in rule may not be. Notably, the Bayesian rule performs particularly well when the values of h are negative. In these cases, the welfare contrasts become smaller, making the assignment problem more challenging. This highlights the robustness of the Bayesian rule to local perturbations that make treatment decisions harder.

To gain further insight into the behavior of the two rules, we visualize the average allocations, $J^{-1}\sum_j \mu_{nh}^{Q,j}$, for $Q = P, E$, under θ_0, $n = 200$, and $\lambda = 0$. Figure 4 shows that the Bayesian rule deviates from the oracle rule only near the decision boundary, while the plug-in rule exhibits substantial deviations even away from it. This relative stability of the Bayesian rule contributes to a sizable risk reduction. A similar, albeit weaker, pattern is observed for $n = 500$.

Figure 3. Comparisons of estimated risks: $n = 500$ (left: smooth welfare, right: directionally differentiable welfare)

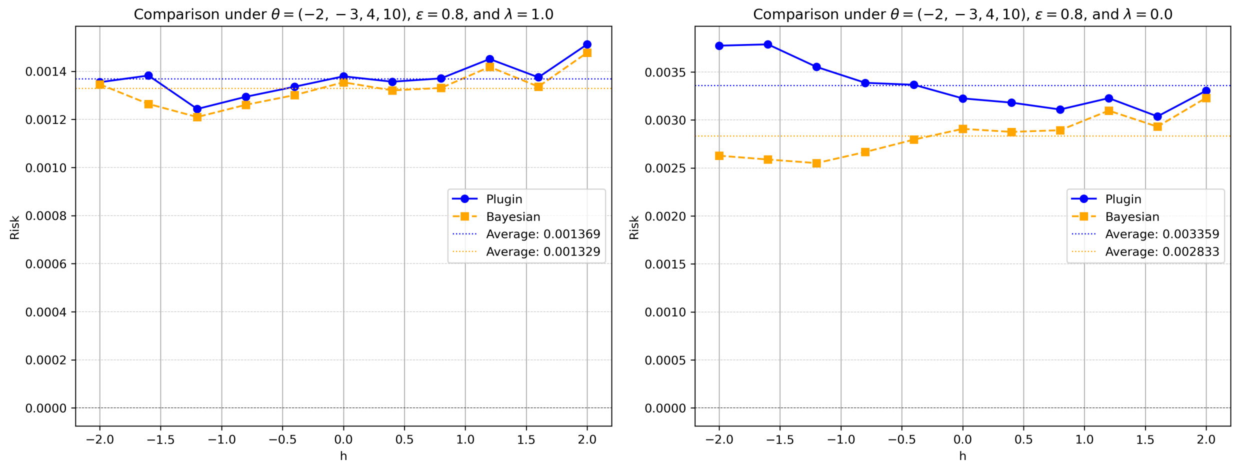


Table 1. Maximum likelihood estimates for math scores

Variables	voucher	age	gender	const	σ^2
Estimates	2.06	-5.5	-0.72	102.77	104.31
Standard error	0.46	0.24	0.44	2.87	5.17

Remark 4.1. Under the current simulation setup, when the total amount of resources is sufficiently small, the kink points of the directionally differentiable welfare do not affect the assignment decision, as the available resources are exhausted before the assignment rule encounters the kink points. In such cases, the behavior of the two rules under directionally differentiable welfare resembles their behavior under the smooth welfare. We set the resource level to 75% to allow interaction between the decision boundary under the oracle rule and the kink points.

5. Empirical application

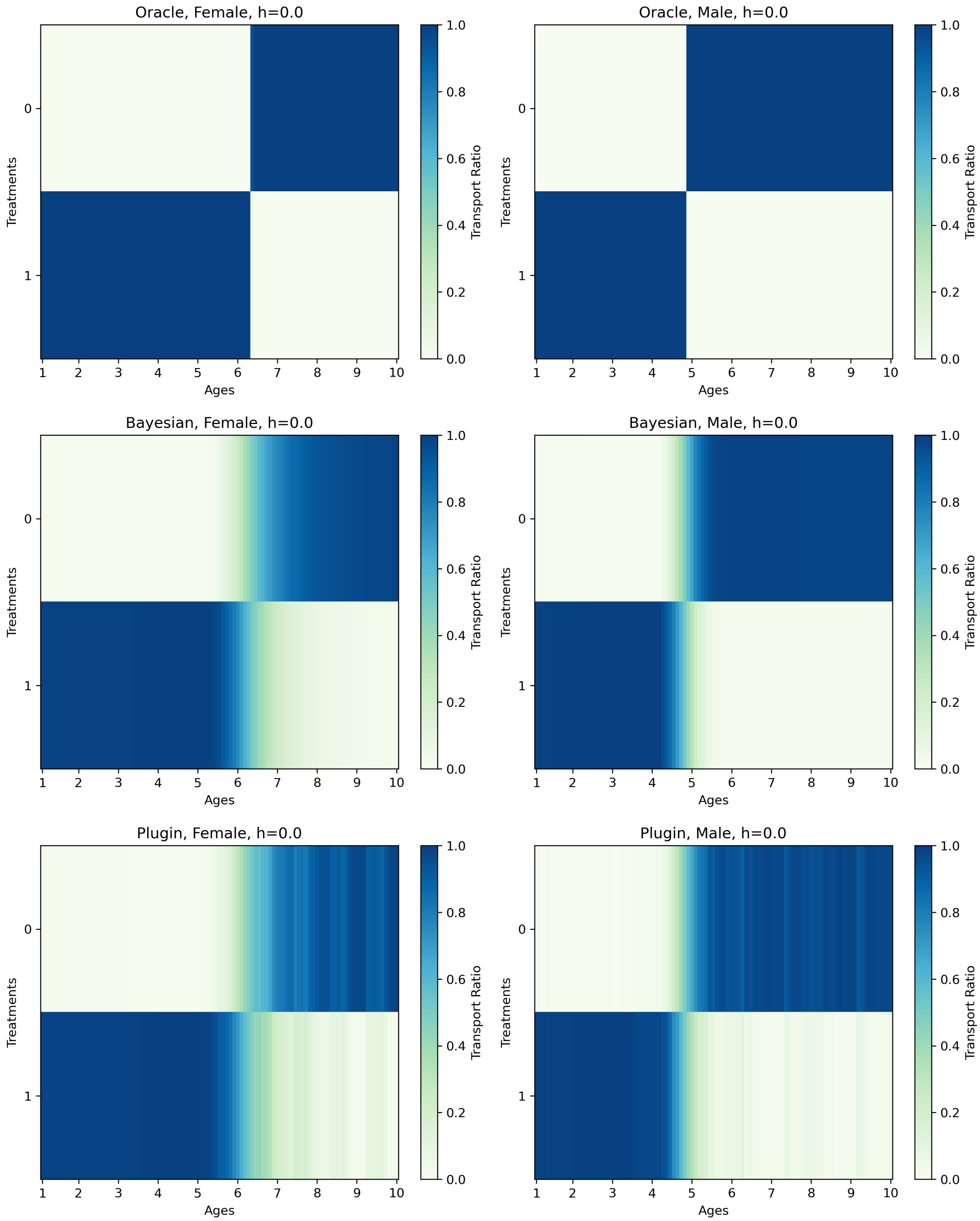
We illustrate our methods using data from Angrist *et al.* (2006) continued from Example 2.3.[11] As outcome variables, Angrist *et al.* (2006) use the test scores in language and math. Since the estimation results are similar between these two, we use math scores as the outcome variable for illustration. We focus on the case where the observed test scores are censored at the tenth percentiles of the test score distribution among test-takers (denoted by τ), in line with the original article, to address selection issues. In addition to the test scores, we observe treatment status, as well as age and sex as covariates. The sample includes 3,541 individuals overall, with 1,788 girls and 1,753 boys. Ages range from 10 to 17 with mean 12.7 and standard deviation 1.3. The maximum likelihood estimates are summarized in Table 1.

We then hypothetically treat the marginal distribution of the covariates in the observed sample as that of the target population, and compute both the plug-in and the Bayesian rules as described

[11]For the replication dataset of the original article, see Angrist *et al.* (2019).

FIGURE 4. Comparison of (average) treatment assignment under directionally differentiable welfare

Note: The upper panels show the oracle rule, the middle show the Bayesian rule, and the lower show the plug-in rule under θ_0 and $n = 200$.



in the previous section. In this example, the planner's utility function is given by:

$$w_R(\theta, x, t, \varepsilon, \lambda) = \lambda w(\theta, x, t) + (1 - \lambda) \max \{w(\theta, x, t) - \varepsilon, \tau\},$$

where

$$w(\theta, x, t) = (x^\top \beta + \alpha t) + (\tau - x^\top \beta - \alpha t)\Phi\left(\frac{\tau - x^\top \beta - \alpha t}{\sigma}\right) + \sigma\phi\left(\frac{\tau - x^\top \beta - \alpha t}{\sigma}\right).$$

Note that this is slightly different from the utility function in the previous section as the outcome variable is censored at $\tau \neq 0$. In what follows, we focus on $\varepsilon = 3.5$ and $\lambda = 0, 1$. We consider the case where we can assign vouchers for 50% of the target population.

Figure 5 shows the allocations under smooth welfare ($\lambda = 1$). As Table 1 shows, age has a negative effect on outcomes. Accordingly, the plug-in rule allocates vouchers to younger individuals. Since the effect of sex is slightly negative, the plug-in rule prioritizes females over males, resulting in the allocation where vouchers are fully allocated to females aged 10-12, while not fully allocated to males at age 12 as the resource is exhausted due to the capacity constraints. In this setting, the Bayesian rule yields exactly the same allocation, which is natural since both rules are optimal under smooth welfare. This also aligns with the simulation result in the previous section: both rules perform similarly when the sample size is large enough.

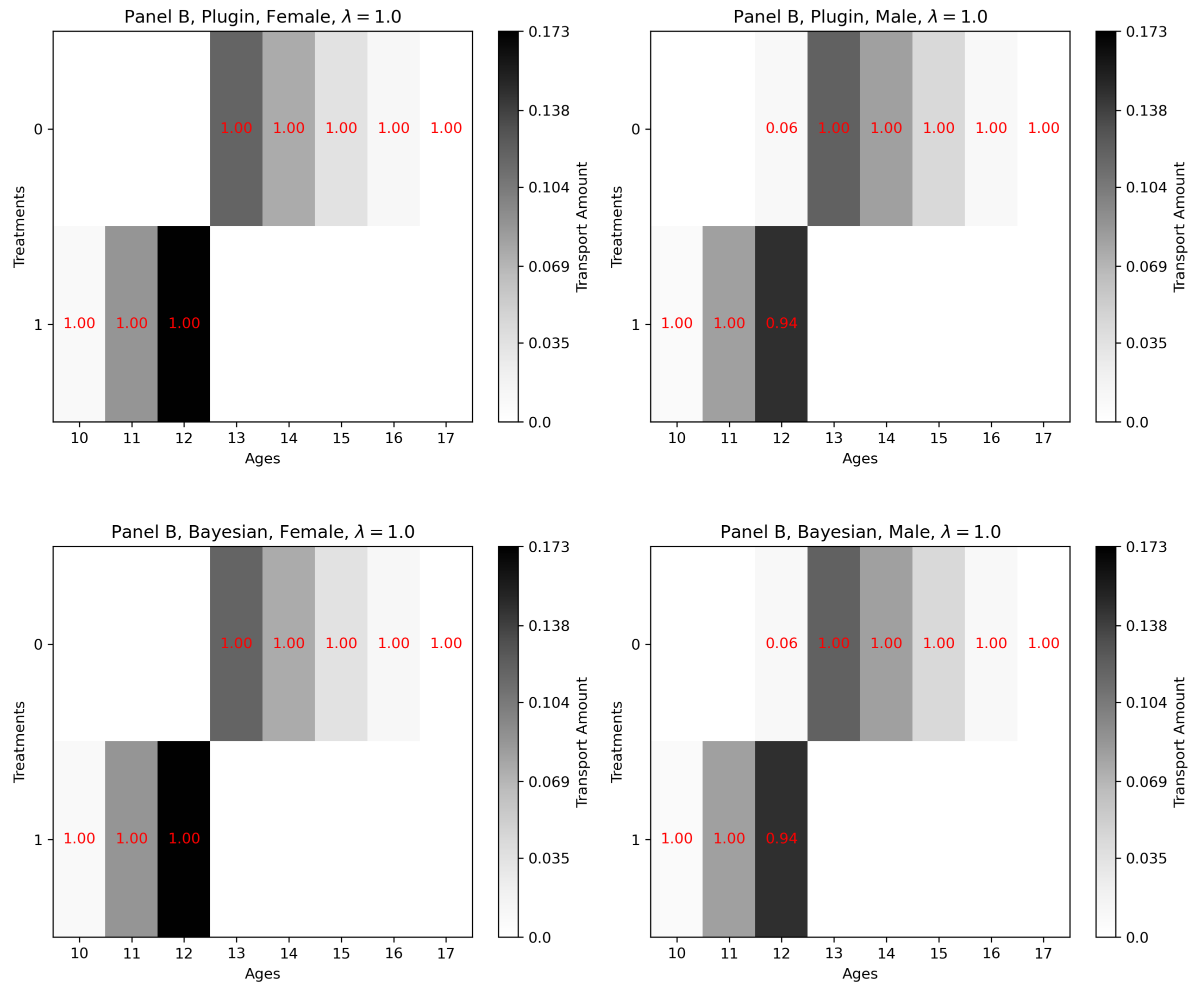
Next, Figure 6 shows the allocations under directionally differentiable welfare ($\lambda = 0$). For the plug-in rule, the value of w_R is censored by τ at age 13 for females and at 12 for males in this setting. As in the previous case, vouchers are fully allocated to females aged 10–12 and males aged 10–11. However, the remaining vouchers are randomly assigned, as the value of w_R is just equal to τ for the rest. For the Bayesian rule, after integrating with respect to the posterior distribution, the value of w_R is censored at τ at age 13 for both females and males. This leads to allocation to males at age 12 until the resource is exhausted, resulting in the same allocation as seen in Figure 5. This illustrates that the plug-in and the Bayesian rules could generate different allocations under $\lambda = 0$.

6. Conclusion

We studied the decision-theoretic optimality of treatment assignment rules under capacity constraints on available treatments. Since such constraints complicate the analysis of optimal rules, we transformed the planner's constrained maximization problem into the unconstrained one using tools from optimal transport theory. This reformulation allows us to search for optimal rules in terms of couplings that automatically satisfy the capacity constraints. We investigated two rules previously studied in the literature—the plug-in rule and the Bayesian rule. Both are average optimal when the planner's utility function is smooth; however, the plug-in rule may no longer be optimal when the planner's utility function is only directionally differentiable. A simulation study supports our theoretical predictions. We demonstrated our methods with a voucher assignment problems for private secondary school attendance using data from Angrist *et al.* (2006).

FIGURE 5. Voucher allocations under smooth welfare

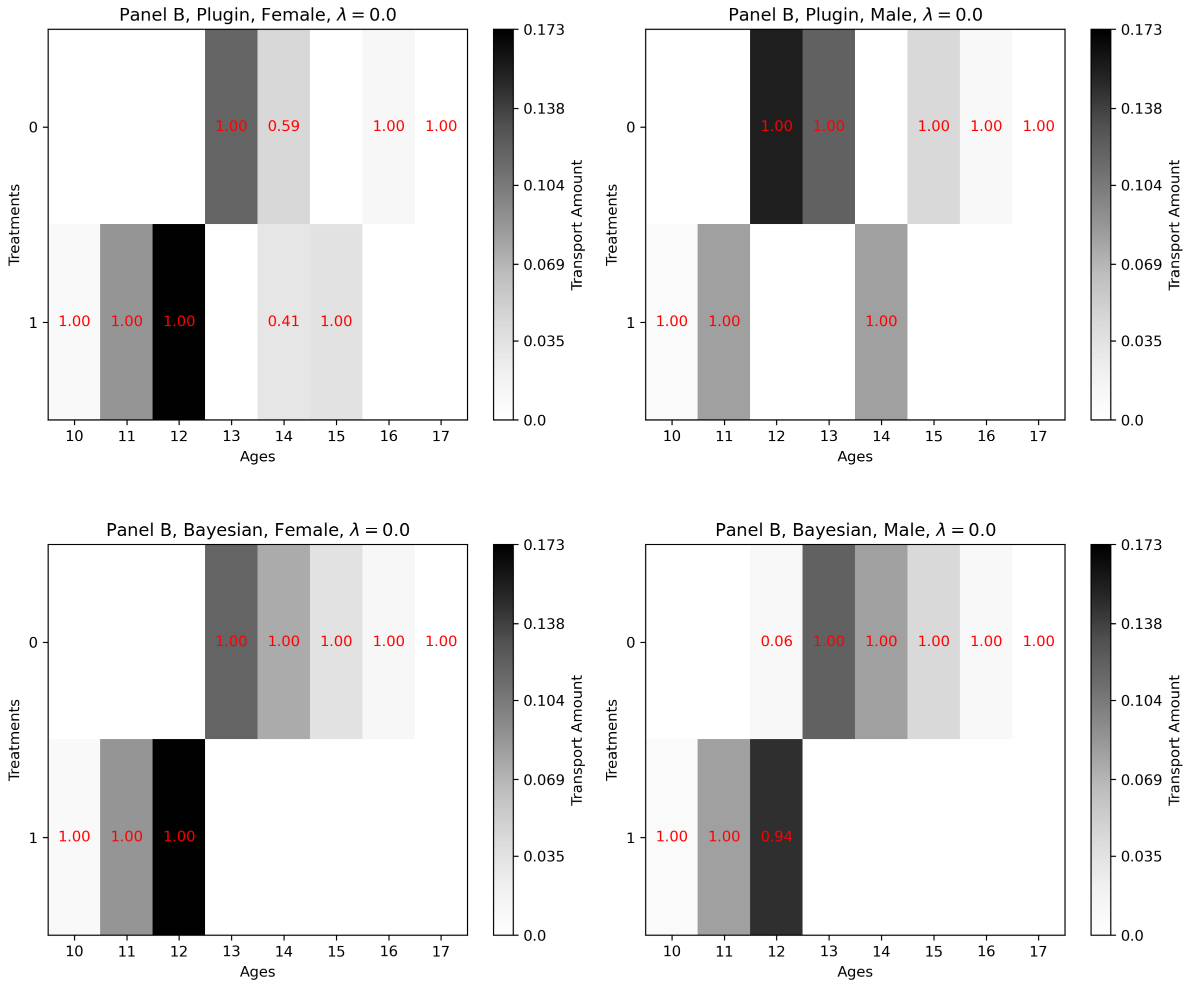
Note: The upper panels show the plug-in rule, and the lower panels show the Bayesian rule. The color intensity represents the density of each cell in F_X, with darker shades indicating higher density.



While we focused on average optimality as the optimality criterion, asymptotic minimax optimality is also a widely used benchmark in local asymptotics frameworks. Kido (2023) provides an asymptotic minimax optimality result when the ATE is partially identified and there are no constraints on available treatments. In that setting, the plug-in rule becomes optimal only when the oracle rule is (Hadamard) differentiable. Given that Hirano and Porter (2009) show the minimax optimality of the plug-in rule under the full differentiability conditions, we conjecture that the modes of differentiability of the planner's utility function w plays a key role in the minimax optimality of the plug-in rule in our setting.

FIGURE 6. Voucher allocations under directionally differentiable welfare

Note: The upper panels show the plug-in rule, and the lower panels show the Bayesian rule. The color intensity represents the density of each cell in F_X, with darker shades indicating higher density.



Appendix A. Proof of Theorem 3.3

A.1. **Preliminaries.** Our proof of Theorem 3.3 proceeds as follows. Lemma A.2 provides sufficient conditions for a rule $\{\mu_n\}$ in $\mathcal{D}$ to be average optimal. Lemma A.3 shows that the Bayesian rule $\{\mu_n^B\}$ satisfies these conditions. Lemmas A.4–A.8 are used to establish Lemma A.3. Additional auxiliary lemmas are relegated to Appendix B.

In what follows, the expectation $\mathbb{E}_{P^n_{\theta_{nh}}}$ can be understood as the outer expectation when μ_n is not measurable. Also, P_h denotes the joint law of $\Delta \sim N(h, I_0^{-1})$ and $U \sim \mathrm{Unif}[0,1]$ in the limit experiment.

The following is a known result and can be found at van der Vaart (1998, Theorem 10.8).

Proposition A.1. *Suppose that model is DQM at θ_0. Let C_n be the ball of radius M_n for a given, arbitrary sequence $M_n \to \infty$. Further, suppose $\int \|\theta\|^p \, \mathrm{d}\pi(\theta) < \infty$. Then, for every measurable function f that grows subpolynomially of order p,*

$$\int f(h)\mathbf{1}_{C_n^c}(h)\pi(\theta_{nh}|Z^n)\mathrm{d}h = o_{P^n_{\theta_0}}(1) \quad \text{as } n \to \infty.$$

A.2. **Proof.** Let $A_0 := \arg\max_{\mu \in \mathcal{M}} W(\theta_0, \mu)$ be the set of couplings that maximize the welfare at θ_0. Note that A_0 is compact by Villani (2009, Corollary 5.21). Define the loss function in the limit experiment by

$$L_\infty(\mu, \Delta) := \int \left[\dot{W}^*_{\mathcal{M},0}[s] - \int \dot{w}_{\theta_0}(x,t;s)\mathrm{d}\mu\right] \mathrm{d}N(\Delta, I_0^{-1})(s),$$

where $\dot{W}^*_{\mathcal{M},0}$ is Hadamard directional derivative of $W^*_{\mathcal{M}} : \Theta \to \mathbb{R}$ at θ_0, which is defined in Lemma B.1.

Lemma A.2. *Let $\{\mu_n\} \in \mathcal{D}$ be any sequence of decision rules that will be matched by μ_∞ in the limit experiment. It holds (i)*

$$\liminf_{n\to\infty} \int \sqrt{n} R(\mu_n, \theta_{nh})\pi(\theta_{nh})\mathrm{d}h \geq \pi(\theta_0) \int_0^1 \int L_\infty(\mu_\infty(\Delta, u), \Delta)\mathrm{d}\Delta \mathrm{d}u,$$

(ii)

$$\limsup_{n\to\infty} \int \sqrt{n} R(\mu_n, \theta_{nh})\pi(\theta_{nh})\mathrm{d}h \leq \pi(\theta_0) \int_0^1 \int L_\infty(\mu_\infty(\Delta, u), \Delta)\mathrm{d}\Delta \mathrm{d}u,$$

and (iii) $\{\mu_n\}$ is average optimal if the matched rule μ_∞ of μ_n satisfies $\mu_\infty(\Delta) \in \arg\min_{\mu \in A_0} L_\infty(\mu, \Delta)$ where $\Delta \sim N(h, I_0^{-1})$.

Proof. (i). By Fatou's lemma,

$$\begin{aligned} &\liminf_{n\to\infty} \int R(\mu_n, \theta_{nh})\pi(\theta_{nh})\mathrm{d}h \\ &\geq \pi(\theta_0) \int \left\{\liminf_{n\to\infty} \mathbb{E}_{P^n_{\theta_{nh}}} \sqrt{n}\left[W^*_{\mathcal{M}}(\theta_{nh}) - W(\theta_{nh}, \mu_n(Z^n))\right]\right\} \mathrm{d}h. \end{aligned} \tag{A.1}$$

For any h,

$$\begin{aligned}
&\mathbb{E}_{P^n_{\theta_{nh}}} \sqrt{n} \left[W^*_{\mathcal{M}}(\theta_{nh}) - W(\theta_{nh}, \mu_n(Z^n))\right] \\
&= \int \sqrt{n} \left(\max_{\mu \in \mathcal{M}} \int w(\theta_{nh}, x, t) \mathrm{d}\mu - \max_{\mu \in \mathcal{M}} \int w(\theta_0, x, t) \mathrm{d}\mu \right) \mathrm{d}P^n_{\theta_{nh}}(z) \\
&\quad + \int \sqrt{n} \left(\max_{\mu \in \mathcal{M}} \int w(\theta_0, x, t) \mathrm{d}\mu - \int w(\theta_0, x, t) \mathrm{d}\mu_n(z) \right) \mathrm{d}P^n_{\theta_{nh}}(z) \\
&\quad - \int \sqrt{n} \left(\int w(\theta_{nh}, x, t) \mathrm{d}\mu_n(z) - \int w(\theta_0, x, t) \mathrm{d}\mu_n(z) \right) \mathrm{d}P^n_{\theta_{nh}}(z).
\end{aligned} \tag{A.2}$$

For the first term of the RHS of (A.2),

$$\begin{aligned}
&\int \sqrt{n} \left(\max_{\mu \in \mathcal{M}} \int w(\theta_{nh}, x, t) \mathrm{d}\mu - \max_{\mu \in \mathcal{M}} \int w(\theta_0, x, t) \mathrm{d}\mu \right) \mathrm{d}P^n_{\theta_{nh}}(z) \\
&= \sqrt{n} \left(\max_{\mu \in \mathcal{M}} \int w(\theta_{nh}, x, t) \mathrm{d}\mu - \max_{\mu \in \mathcal{M}} \int w(\theta_0, x, t) \mathrm{d}\mu \right) \to \dot{W}^*_{\mathcal{M},0}[h],
\end{aligned}$$

where the convergence follows from Lemma B.1. For the second term of the RHS of (A.2), it is clear that

$$\int \sqrt{n} \left(\max_{\mu \in \mathcal{M}} \int w(\theta_0, x, t) \mathrm{d}\mu - \int w(\theta_0, x, t) \mathrm{d}\mu_n(z) \right) \mathrm{d}P^n_{\theta_{nh}}(z) \geq 0.$$

For the third term of the RHS of (A.2),

$$\begin{aligned}
&\int \sqrt{n} \left(\int w(\theta_{nh}, x, t) \mathrm{d}\mu_n(z) - \int w(\theta_0, x, t) \mathrm{d}\mu_n(z) \right) \mathrm{d}P^n_{\theta_{nh}}(z) \\
&= \int \int \dot{w}_{\theta_0}(x, t; h) \mathrm{d}\mu_n(z) \mathrm{d}P^n_{\theta_{nh}}(z) \\
&\quad + \int \int \left\{ \sqrt{n} \left[w(\theta_{nh}, x, t) - w(\theta_0, x, t)\right] - \dot{w}_{\theta_0}(x, t; h) \right\} \mathrm{d}\mu_n(z) \mathrm{d}P^n_{\theta_{nh}}(z)
\end{aligned}$$

where $\dot{w}_{\theta_0}(x, t; h)$ is the directional derivative of $w(\theta, x, t)$ at θ_0. By Assumption 3.4 (i),

$$\begin{aligned}
&\left| \int \int \left\{ \sqrt{n} \left[w(\theta_{nh}, x, t) - w(\theta_0, x, t)\right] - \dot{w}_{\theta_0}(x, t; h) \right\} \mathrm{d}\mu_n(z) \mathrm{d}P^n_{\theta_{nh}}(z) \right| \\
&\leq \max_{(x,t) \in \mathcal{X} \times \mathcal{T}} \left| \sqrt{n} \left[w(\theta_{nh}, x, t) - w(\theta_0, x, t)\right] - \dot{w}_{\theta_0}(x, t; h) \right| \to 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\liminf_{n \to \infty} \mathbb{E}_{P^n_{\theta_{nh}}} \sqrt{n} \left[W^*_{\mathcal{M}}(\theta_{nh}) - W(\theta_{nh}, \mu_n(Z^n))\right] \\
&\geq \dot{W}^*_{\mathcal{M},0}[h] - \limsup_{n \to \infty} \int \int \dot{w}_{\theta_0}(x, t; h) \mathrm{d}\mu_n(z) \mathrm{d}P^n_{\theta_{nh}}(z) \\
&\equiv \dot{W}^*_{\mathcal{M},0}[h] - \limsup_{n \to \infty} \int \varphi(\mu_n(z)) \mathrm{d}P^n_{\theta_{nh}}(z) \\
&\geq \dot{W}^*_{\mathcal{M},0}[h] - \mathbb{E}_{(\Delta, U) \sim P_h} \varphi(\mu_\infty(\Delta, U))
\end{aligned}$$

from the portmanteau theorem in metric spaces since the map $\varphi : \mu \mapsto \int \dot{w}_{\theta_0}(x,t;h)\mathrm{d}\mu$ is bounded continuous on $\mathcal{M}$. Thus, the RHS of (A.1) is bounded below by

$$\begin{aligned}
&\pi(\theta_0) \int \liminf_{n\to\infty} \mathbb{E}_{P^n_{\theta_{nh}}} \sqrt{n}\left[W^*_{\mathcal{M}}(\theta_{nh}) - W(\theta_{nh}, \mu_n(Z^n))\right] \mathrm{d}h \\
&\geq \pi(\theta_0) \int \mathbb{E}_{P_h}\left[\dot{W}^*_{\mathcal{M},0}[h] - \varphi(\mu_\infty)\right] \mathrm{d}h \\
&= \pi(\theta_0) \int \int_0^1 \int \left[\dot{W}^*_{\mathcal{M},0}[h] - \varphi(\mu_\infty(\Delta,u))\right] \mathrm{d}N(h, I_0^{-1})(\Delta)\mathrm{d}u\mathrm{d}h \\
&= \pi(\theta_0) \int_0^1 \int \int \left[\dot{W}^*_{\mathcal{M},0}[s] - \int \dot{w}_{\theta_0}(x,t;s)\mathrm{d}\mu_\infty(\Delta,u)\right] \mathrm{d}N(\Delta, I_0^{-1})(s)\mathrm{d}\Delta\mathrm{d}u,
\end{aligned}$$

where the last equality follows by Tonelli's theorem since the integrand is nonnegative. By the definition of L_∞, the last display is equal to

$$\int_0^1 \int L_\infty(\mu_\infty(\Delta,u),\Delta)\mathrm{d}\Delta\mathrm{d}u.$$

It should be noted that $L_\infty(\mu_\infty(\Delta,u),\Delta)$ can depend on u only through μ_∞.

(ii). If we admit applying the reverse Fatou lemma, we obtain

$$\limsup_{n\to\infty} \int \sqrt{n} R(\mu_n, \theta_{nh})\pi(\theta_{nh})\mathrm{d}h \leq \pi(\theta_0) \int \left\{ \limsup_{n\to\infty} \mathbb{E}_{P^n_{\theta_{nh}}} \sqrt{n}\left[W^*_{\mathcal{M}}(\theta_{nh}) - W(\theta_{nh}, \mu_n(Z^n))\right] \right\} \mathrm{d}h.$$

The argument can be carried out analogously to (i). However, for the second term of the RHS of (A.2),

$$\begin{aligned}
&\int \sqrt{n}\left(\max_{\mu\in\mathcal{M}} \int w(\theta_0,x,t)\mathrm{d}\mu - \int w(\theta_0,x,t)\mathrm{d}\mu_n(z)\right) \mathrm{d}P^n_{\theta_{nh}}(z) \\
&= \int_{\{\mu_n\in A_0\}} \sqrt{n}\left(\max_{\mu\in\mathcal{M}} \int w(\theta_0,x,t)\mathrm{d}\mu - \int w(\theta_0,x,t)\mathrm{d}\mu_n(z)\right) \mathrm{d}P^n_{\theta_{nh}}(z) \\
&\quad + \int_{\{\mu_n\notin A_0\}} \sqrt{n}\left(\max_{\mu\in\mathcal{M}} \int w(\theta_0,x,t)\mathrm{d}\mu - \int w(\theta_0,x,t)\mathrm{d}\mu_n(z)\right) \mathrm{d}P^n_{\theta_{nh}}(z) \\
&= \int_{\{\mu_n\notin A_0\}} \sqrt{n}\left(\max_{\mu\in\mathcal{M}} \int w(\theta_0,x,t)\mathrm{d}\mu - \int w(\theta_0,x,t)\mathrm{d}\mu_n(z)\right) \mathrm{d}P^n_{\theta_{nh}}(z),
\end{aligned}$$

where the second equality follows from $\max_{\mu\in\mathcal{M}} \int w(\theta_0,x,t)\mathrm{d}\mu = \max_{\mu\in A_0} \int w(\theta_0,x,t)\mathrm{d}\mu$. Further,

$$\begin{aligned}
&\int_{\{\mu_n\notin A_0\}} \sqrt{n}\left(\max_{\mu\in\mathcal{M}} \int w(\theta_0,x,t)\mathrm{d}\mu - \int w(\theta_0,x,t)\mathrm{d}\mu_n(z)\right) \mathrm{d}P^n_{\theta_{nh}}(z) \\
&\leq \sqrt{n} P^n_{\theta_{nh}}\left(\mu_n \notin A_0\right) \times 2 \max_{(x,t)\in\mathcal{X}\times\mathcal{T}} |w(\theta_0,x,t)| \to 0,
\end{aligned}$$

as $n \to \infty$ from the definition of $\{\mu_n\} \in \mathcal{D}$. Thus, applying the portmanteau theorem yields

$$\limsup_{n\to\infty} \mathbb{E}_{P^n_{\theta_{nh}}} \sqrt{n}\left[W^*_{\mathcal{M}}(\theta_{nh}) - W(\theta_{nh}, \mu_n(Z^n))\right] \leq \dot{W}^*_{\mathcal{M},0}[h] - \mathbb{E}_{P_h}\left[\int \dot{w}_{\theta_0}(x,t;h)\mathrm{d}\mu_\infty\right]. \tag{A.3}$$

Then we can conclude similarly as in (i).

Finally, we check the validity of applying the reverse Fatou lemma. Let

$$g_n(h) := \mathbb{E}_{P^n_{\theta_{nh}}} \sqrt{n} \left[W^*_{\mathcal{M}}(\theta_{nh}) - W(\theta_{nh}, \mu_n(Z^n)) \right].$$

From (A.3), there exists N such that for all $n > N$,

$$g_n(h) \leq \dot{W}^*_{\mathcal{M},0}[h] - \mathbb{E}_{P_h} \left[\int \dot{w}_{\theta_0}(x, t; h) \mathrm{d}\mu_\infty \right] + \varepsilon.$$

Note that $g_n(h)$ is bounded for all $n \leq N$. Then define

$$g(h) := \max \left\{ g_1(h), \ldots, g_N(h), \dot{W}^*_{\mathcal{M},0}[h] - \mathbb{E}_{P_h} \left[\int \dot{w}_{\theta_0}(x, t; h) \mathrm{d}\mu_\infty \right] + \varepsilon \right\}.$$

Thus we have $g_n(h)\pi(\theta_{nh}) \leq \sup_{\theta \in \Theta} \pi(\theta) \times g(h)$ for all $n \in \mathbb{N}$, with $\sup_{\theta \in \Theta} \pi(\theta) < \infty$ from Assumption 3.5 (i).

(iii). Fix any sequence of rules $\{\mu'_n\} \in \mathcal{D}$, and let μ'_∞ be the matched rule of μ'_n. Combining (i) and (ii) yields

$$\begin{aligned}
\liminf_{n \to \infty} \int \sqrt{n} R(\mu_n, \theta_{nh}) \pi(\theta_{nh}) \mathrm{d}h &\leq \limsup_{n \to \infty} \int \sqrt{n} R(\mu_n, \theta_{nh}) \pi(\theta_{nh}) \mathrm{d}h \\
&\leq \pi(\theta_0) \int_0^1 \int L_\infty(\mu_\infty(\Delta, u), \Delta) \mathrm{d}\Delta \mathrm{d}u \\
&\leq \pi(\theta_0) \int_0^1 \int L_\infty(\mu'_\infty(\Delta, u), \Delta) \mathrm{d}\Delta \mathrm{d}u \\
&\leq \liminf_{n \to \infty} \int \sqrt{n} R(\mu'_n, \theta_{nh}) \pi(\theta_{nh}) \mathrm{d}h,
\end{aligned}$$

which completes the proof. □

Let $(\mathbb{D}, \|\cdot\|_{\mathbb{D}})$ be the product metric space induced by $(\mathcal{M}, d_W)$ and $([0,1]. |\cdot|)$. Let $\ell^\infty(\mathbb{D}) := \{f : \mathbb{D} \to \mathbb{R} : \sup_{(\mu,\varepsilon) \in \mathbb{D}} |f(\mu, \varepsilon)| < \infty\}$. Define

$$\begin{aligned}
(\mu, \varepsilon) \mapsto & \mathcal{Q}_n(\mu, \varepsilon; z) := \int \left[\int \sqrt{n} \left(w(\theta_{nh}, x, t) - w(\theta_0, x, t) \right) \mathrm{d}\mu \right] \pi_n(\theta_{nh} | z) \mathrm{d}h - \varepsilon H(\mu), \\
(\mu, \varepsilon) \mapsto & Q_\infty(\mu, \varepsilon; \Delta) := \int \left[\int \dot{w}_{\theta_0}(x, t; h) \mathrm{d}\mu \right] \mathrm{d}N(\Delta, I_0^{-1})(h) - \varepsilon H(\mu).
\end{aligned}$$

In the same manner as the definition of $\mu^B_n(Z^n)$, we define $\mu^*_\infty(\Delta)$ as the limit of $\mu^*_{\infty,\varepsilon}(\Delta)$ where

$$\mu^*_{\infty,\varepsilon}(\Delta) := \arg\max_{\mu \in A_0} \int \int \dot{w}_{\theta_0}(x, t; s) \mathrm{d}N(\Delta, I_0^{-1})(s) \mathrm{d}\mu - \varepsilon H(\mu).$$

It is easy to see that $\mu^*_\infty(\Delta) \in \arg\min_{\mu \in A_0} L_\infty(\mu, \Delta)$.

Lemma A.3. *The Bayesian rule* $\{\mu^B_n(Z^n)\}$ *satisfies* $\sqrt{n} P^n_{\theta_{nh}} \left(\mu^B_n(Z^n) \notin A_0 \right) \to 0$ *and (ii)* $\mu^B_n \overset{h}{\rightsquigarrow} \mu^*_\infty$ *as* $n \to \infty$.

Proof. (i). We claim that for any $\theta_0 \in \Theta$, there are $\bar{n}$ and $\varepsilon'_n(\bar{n})$ (which is at the order of $n^{\alpha+1}$ for some $\alpha \geq 1$) such that for all $n \geq \bar{n}$,

$$P^n_{\theta_{nh}} \left(\mu^B_n(z) \notin A_0 \right) \leq P^n_{\theta_{nh}} \left(\pi_n \left(N_{1/n}(\theta_0)^c | z \right) > 2\varepsilon'_n \right),$$

where $N_\varepsilon(\theta_0) := \{\theta : \|\theta - \theta_0\| < \varepsilon\}$ for $\varepsilon > 0$, and $\pi_n(A|z) := \int_A \pi_n(\theta|z)\mathrm{d}\theta$.[12] Then the conclusion follows since Christensen *et al.* (2025, Lemmas 9–11) imply that $\sqrt{n}P^n_{\theta_{nh}}\left(\pi_n\left(N_{1/n}(\theta_0)^c|z\right) > 2\varepsilon'_n\right) \to 0$ as $n \to \infty$.

It is sufficient to show that for any z,

$$\pi_n\left(N_{1/n}(\theta_0)^c|z\right) \le 2\varepsilon'_n \implies \mu^B_n(z) \in A_0. \tag{A.4}$$

It is trivial if $A_0 = \mathcal{M}$, so suppose $A_0 \subsetneq \mathcal{M}$. By Lemma A.4 below, there exists $\bar{n}$ such that $n \ge \bar{n}$ implies

$$\min_{\mu\in A_0}\int_{N_{1/n}(\theta_0)} W(\theta,\mu)\pi_n(\theta|z)\mathrm{d}\theta > \sup_{\nu\notin A_0}\int_{N_{1/n}(\theta_0)} W(\theta,\nu)\pi_n(\theta|z)\mathrm{d}\theta$$

Then there exists $\alpha = \alpha(\bar{n}) \ge 1$ such that

$$\min_{\mu\in A_0}\int_{N_{1/n}(\theta_0)} W(\theta,\mu)\pi_n(\theta|z)\mathrm{d}\theta > \sup_{\nu\notin A_0}\int_{N_{1/n}(\theta_0)} W(\theta,\nu)\pi_n(\theta|z)\mathrm{d}\theta + \frac{1}{\bar{n}^\alpha},$$

which implies that for $n \ge \bar{n}$,

$$\min_{\mu\in A_0}\int_{N_{1/n}(\theta_0)} W(\theta,\mu)\pi_n(\theta|z)\mathrm{d}\theta > \sup_{\nu\notin A_0}\int_{N_{1/n}(\theta_0)} W(\theta,\nu)\pi_n(\theta|z)\mathrm{d}\theta + \frac{1}{n^\alpha}.$$

Thus, we have

$$\begin{aligned}
\min_{\mu\in A_0}\left(\int W(\theta,\mu)\pi_n(\theta|z)\mathrm{d}\theta\right) &= \min_{\mu\in A_0}\left(\int_{N_{1/n}(\theta_0)} W(\theta,\mu)\pi_n(\theta|z)\mathrm{d}\theta + \int_{N_{1/n}(\theta_0)^c} W(\theta,\mu)\pi_n(\theta|z)\mathrm{d}\theta\right)\\
&\ge \min_{\mu\in A_0}\int_{N_{1/n}(\theta_0)} W(\theta,\mu)\pi_n(\theta|z)\mathrm{d}\theta + \min_{\mu\in A_0}\int_{N_{1/n}(\theta_0)^c} W(\theta,\mu)\pi_n(\theta|z)\mathrm{d}\theta\\
&\ge \min_{\mu\in A_0}\int_{N_{1/n}(\theta_0)} W(\theta,\mu)\pi_n(\theta|z)\mathrm{d}\theta - \pi_n\left(N_{1/n}(\theta_0)^c|z\right)M,
\end{aligned}$$

where $\sup_{\theta,\mu}|W(\theta,\mu)| \le M < \infty$. Also,

$$\sup_{\nu\notin A_0}\int W(\theta,\nu)\pi_n(\theta|z)\mathrm{d}\theta \le \sup_{\nu\notin A_0}\int_{N_{1/n}(\theta_0)} W(\theta,\nu)\pi_n(\theta|z)\mathrm{d}\theta + \pi_n\left(N_{1/n}(\theta_0)^c|z\right)M.$$

For the promise for (A.4), we choose $\varepsilon'_n > 0$ that satisfies $\left(n^{\alpha+1}2M\right)^{-1} \le 2\varepsilon'_n < (n^\alpha 2M)^{-1}$, which leads to

$$\pi_n\left(N_\varepsilon(\theta_0)^c|z\right) < \frac{1}{n^\alpha 2M}.$$

Then it follows that

$$\begin{aligned}
&\min_{\mu\in A_0}\int W(\theta,\mu)\pi_n(\theta|z)\mathrm{d}\theta - \sup_{\nu\notin A_0}\int W(\theta,\nu)\pi_n(\theta|z)\mathrm{d}\theta\\
\ge&\min_{\mu\in A_0}\int_{N_{1/n}(\theta_0)} W(\theta,\mu)\pi_n(\theta|z)\mathrm{d}\theta - \sup_{\nu\notin A_0}\int_{N_{1/n}(\theta_0)} W(\theta,\nu)\pi_n(\theta|z)\mathrm{d}\theta - 2\pi_n\left(N_{1/n}(\theta_0)^c|z\right)M\\
\ge&\frac{1}{n^\alpha} - 2\pi_n\left(N_{1/n}(\theta_0)^c|z\right)M > 0,
\end{aligned}$$

[12] This statement is an adaptation of Christensen *et al.* (2025, Lemma 8). Their proof cannot directly apply to our setting because their arguments could fail when the set of actions, $\mathcal{M}$ in our notation, is not finite.

which implies

$$\min_{\mu \in A_0} \int W(\theta, \mu)\pi_n(\theta|z)\mathrm{d}\theta > \sup_{\nu \notin A_0} \int W(\theta, \nu)\pi_n(\theta|z)\mathrm{d}\theta.$$

Thus we conclude $\mu_n^B(z) \in A_0$.

(ii). From the first statement, it follows that the asymptotic distribution of $\mu_n^B(Z^n)$ has the support only on A_0. Hence, for sufficiently large n, μ_n^B equivalently solves

$$\mu_n^B(z) \in \arg\max_{\mu \in A_0} \mathcal{Q}_n(\mu, 0; z).$$

By the definition of μ_n^B,

$$\mathcal{Q}_n(\mu_n^B(z), 0; z) = \lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon; z).$$

Take any closed subset G of $\mathcal{M}$. Note that this closedness is in terms of $(\mathcal{M}, d_W)$. By the Portmanteau lemma, it is sufficient to show that

$$\limsup_{n \to \infty} P^n_{\theta_{nh}} \left(\mu_n^B(z) \in G\right) \leq P_h \left(\mu^*_\infty(\Delta) \in G\right)$$

for the conclusion.

By Lemma A.6, for each n,

$$\left\{\mu_n^B(z) \in G\right\} = \left\{\lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0 \cap G} \mathcal{Q}_n(\mu, \varepsilon; z) = \lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon; z)\right\}.$$

By Lemmas A.7 and A.8, it follows that $\mathcal{Q}_n \overset{h}{\rightsquigarrow} Q_\infty$ in $\mathcal{F}$ as $n \to \infty$, where

$$\mathcal{F} = \left\{f \in \ell^\infty(\mathbb{D}) : \lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0} f(\mu, \varepsilon) \text{ exists}\right\}.$$

By Lemma B.3, the operator $f \mapsto \lim_{\varepsilon \downarrow 0} \max_{\mu \in S} f(\mu, \varepsilon)$ is continuous for any closed $S \subset A_0$ at any $f \in \mathcal{F}$. Applying the continuous mapping theorem yields

$$\lim_{\varepsilon \downarrow 0} \max_{\mu \in S} \mathcal{Q}_n(\mu, \varepsilon; z) \overset{h}{\rightsquigarrow} \lim_{\varepsilon \downarrow 0} \max_{\mu \in S} Q_\infty(\mu, \varepsilon; \Delta) \quad \text{as } n \to \infty.$$

Then

$$\begin{aligned}
&\limsup_{n \to \infty} P^n_{\theta_{nh}} \left(\mu_n^B(z) \in G\right) \\
=&\limsup_{n \to \infty} P^n_{\theta_{nh}} \left(\lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0 \cap G} \mathcal{Q}_n(\mu, \varepsilon; z) = \lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon; z)\right) \\
\leq&P_h \left(\lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0 \cap G} Q_\infty(\mu, \varepsilon; \Delta) = \lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0} Q_\infty(\mu, \varepsilon; \Delta)\right),
\end{aligned}$$

where the inequality follows from the Portmanteau lemma. Since we can derive the equivalence of the events

$$\{\mu^*_\infty(\Delta) \in G\} = \left\{\lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0 \cap G} Q_\infty(\mu, \varepsilon; \Delta) = \lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0} Q_\infty(\mu, \varepsilon; \Delta)\right\}$$

in the same manner, applying the Portmanteau lemma again yields $\mu_n^B \overset{h}{\rightsquigarrow} \mu^*_\infty$ as $n \to \infty$. □

Lemma A.4. *There exists $\bar{n} > 0$ such that for all $n \geq \bar{n}$,*

$$\min_{\mu \in A_0} \int_{N_{1/n}(\theta_0)} W(\theta, \mu)\pi_n(\theta|z)\mathrm{d}\theta > \sup_{\nu \notin A_0} \int_{N_{1/n}(\theta_0)} W(\theta, \nu)\pi_n(\theta|z)\mathrm{d}\theta. \tag{A.5}$$

Proof. Let $g_n(\mu) := \int_{N_{1/n}(\theta_0)} W(\theta, \mu)\pi_n(\theta)\mathrm{d}\theta$ and $V_{1/n} := \int_{N_{1/n}(\theta_0)} \pi_n(\theta)\mathrm{d}\theta$. First, we claim that $V_{1/n}^{-1} g_n(\mu)$ converges to $W(\theta_0, \mu)$ uniformly over $\mathcal{M}$ as $n \to \infty$; i.e., for all $\eta > 0$, there exists n_η such that for all $\mu \in \mathcal{M}$,

$$n \geq n_\eta \implies \left| V_{1/n}^{-1} g_n(\mu) - W(\theta_0, \mu) \right| < \eta.$$

To show this claim, we argue that (i) for each $\mu \in \mathcal{M}$, $V_{1/n}^{-1} g_n(\mu)$ converges to $W(\theta_0, \mu)$ in pointwise, and (ii) $\{g_n\}_{n \in \mathbb{N}}$ is equicontinuous; i.e., for all $\eta > 0$ and all $\mu \in \mathcal{M}$, there exists $\delta_{(\eta,\mu)} > 0$ such that for all $n \in \mathbb{N}$ and all $\nu \in \mathcal{M}$,

$$d_W(\mu, \nu) < \delta_{(\eta,\mu)} \implies |g_n(\mu) - g_n(\nu)| < \eta.$$

Combining with the compactness of $\mathcal{M}$, the uniform convergence follows from these two.

To see (i), note that

$$\left| V_{1/n}^{-1} g_n(\mu) - W(\theta_0, \mu) \right| \leq V_{1/n}^{-1} \int_{N_{1/n}(\theta_0)} |W(\theta, \mu) - W(\theta_0, \mu)| \, \pi_n(\theta|z)\mathrm{d}\theta. \tag{A.6}$$

Fix $\eta > 0$. Since the map $\theta \mapsto W(\theta, \mu)$ is continuous at θ_0, there exists $\delta > 0$ such that

$$\theta \in N_\delta(\theta_0) \implies |W(\theta, \mu) - W(\theta_0, \mu)| < \eta.$$

Then for all n with $n^{-1} < \delta$, RHS of (A.6) is bounded above by η.

To see (ii), fix $\eta > 0$ and $\mu \in \mathcal{M}$. First, note that $g_n(\mu)$ can be written as

$$g_n(\mu) = \mathbb{E}_\mu[\Psi_n(x, t)], \quad \Psi_n(x, t) := \int_{N_{1/n}(\theta_0)} w(\theta, x, t)\pi_n(\theta|z)\mathrm{d}\theta,$$

where $\mathbb{E}_\mu$ denotes the expectation with respect to the coupling μ. Note that $(x, t) \mapsto w(\theta, x, t)$ is uniformly continuous and bounded since $\mathcal{X} \times \mathcal{T}$ is compact. Then a function $(x, t) \mapsto w_k(\theta, x, t)$ defined by

$$w_k(\theta, x, t) := \inf \left\{ w(\theta, x', t') + kd((x, t), (x', t')) : (x', t') \in \mathcal{X} \times \mathcal{T} \right\}$$

is k-Lipschitz continuous, and converges uniformly to $w(\theta, x, t)$ from below as $k \to \infty$ (see e.g., Heinonen (2001, Theorem 6.8)). Lemma A.5 below further extends that the convergence holds uniformly over Θ; i.e., for all $\eta > 0$, there is a sufficiently large $K = K(\eta)$ such that

$$\sup_{\theta \in \Theta} \max_{(x,t) \in \mathcal{X} \times \mathcal{T}} |w(\theta, x, t) - w_K(\theta, x, t)| < \frac{\eta}{3}.$$

Given this K, define

$$\Psi_n^K(x, t) := \int_{N_{1/n}(\theta_0)} w_K(\theta, x, t)\pi_n(\theta|z)\mathrm{d}\theta.$$

Then $\Psi_n^K(x,t)$ is also Lipschitz continuous whose Lipschitz constant is less than or equal to K. Therefore,

$$\begin{aligned}|g_n(\mu)-g_n(\nu)| \leq &\mathbb{E}_\mu\left|\Psi_n(x,t)-\Psi_n^K(x,t)\right| + \left|\mathbb{E}_\mu[\Psi_n^K(x,t)]-\mathbb{E}_\nu[\Psi_n^K(x,t)]\right| + \mathbb{E}_\nu\left|\Psi_n^K(x,t)-\Psi_n(x,t)\right| \\ <&\frac{2}{3}\eta V_{1/n} + K d_W(\mu,\nu) \leq \frac{2}{3}\eta + K d_W(\mu,\nu),\end{aligned}$$

where the second inequality follows from the Kantorovich-Rubinstein duality (Villani, 2009, Theorem 5.10). Thus, we obtain

$$d_W(\mu,\nu) < \frac{\eta}{3K} \implies |g_n(\mu)-g_n(\nu)| < \eta.$$

Therefore $\{g_n\}_{n\in\mathbb{N}}$ is an equicontinuous family.

Finally, we show (A.5). By Assumption 3.6, there exists $\eta > 0$ such that for all $\mu \in A_0$,

$$W(\theta_0,\mu) > \sup_{\nu\notin A_0} W(\theta_0,\nu) + \eta. \tag{A.7}$$

By the uniform convergence shown above, there exists n_η such that for all $\mu \in \mathcal{M}$,

$$n \geq n_\eta \implies \left|V_{1/n}^{-1} g_n(\mu) - W(\theta_0,\mu)\right| < \frac{\eta}{3}.$$

Then fix $n \geq n_\eta$, and let $\mu_n \in \arg\min_{\mu\in A_0} g_n(\mu)$. For any $\nu \notin A_0$, we obtain

$$\begin{aligned}\min_{\mu\in A_0} g_n(\mu) &= g_n(\mu_n) \\ &> V_{1/n}\left[W(\theta_0,\mu_n) - \frac{\eta}{3}\right] \\ &> V_{1/n}\left[W(\theta_0,\nu) + \eta - \frac{\eta}{3}\right] \\ &> g_n(\nu) + \frac{\eta}{3},\end{aligned}$$

where the second inequality follows from (A.7). Thus we obtain

$$\min_{\mu\in A_0} g_n(\mu) \geq \left\{\sup_{\nu\notin A_0} g_n(\nu)\right\} + \frac{\eta}{3}.$$

Since $\eta > 0$ does not depend on $\nu \notin A_0$, we conclude $n \geq n_\eta$ implies $\min_{\mu\in A_0} g_n(\mu) > \sup_{\nu\notin A_0} g_n(\nu)$. □

Lemma A.5. *For all $\eta > 0$, there is a sufficiently large $K = K(\eta)$ such that*

$$\sup_{\theta\in\Theta}\max_{(x,t)\in\mathcal{X}\times\mathcal{T}} |w(\theta,x,t) - w_K(\theta,x,t)| < \eta.$$

Proof. To simplify the notation, let $\mathcal{Y} = \mathcal{X}\times\mathcal{T}$. Let D be the diameter of $\mathcal{Y}$. Since $y \mapsto w(\theta,y)$ is continuous uniformly over Θ (Assumption 3.3 (ii)) and $\mathcal{Y}$ is compact (Assumption 3.2), we have

$$\Omega(\delta) := \sup_{\theta\in\Theta}\sup_{d(y,y')<\delta}\left|w(\theta,y) - w(\theta,y')\right| \to 0 \quad \text{as } \delta \downarrow 0.$$

Fix any $\theta \in \Theta$ and $y \in \mathcal{Y}$. Then for all $y' \in \mathcal{Y}$,

$$w(\theta, y') \geq w(\theta, y) - \left|w(\theta, y) - w(\theta, y')\right| \geq w(\theta, y) - \Omega(d(y, y')),$$

where the last inequality follows from the definition of Ω. Adding $kd(y, y')$ to both sides and taking the infimum with respect to y' yields

$$w_k(\theta, y) \geq w(\theta, y) + \inf_{r \in [0,D]} \{kr - \Omega(r)\},$$

which implies

$$w(\theta, y) - w_k(\theta, y) \leq \sup_{r \in [0,D]} \phi_k(r), \quad \phi_k(r) := \Omega(r) - kr.$$

By the definition of Ω, there exists $\delta > 0$ such that $\Omega(\delta) < \eta$. Note that Ω is non-decreasing function. Then if $0 \leq r < \delta$, we have $\phi_k(r) \leq \Omega(\delta) < \eta$. If $r \geq \delta$, we have $\phi_k(r) \leq \Omega(D) - k\delta$. Thus $\sup_{r \in [0,D]} \phi_k(r) \leq \{\Omega(\delta), \Omega(D) - k\delta\}$. Hence, for sufficiently large K, it follows $\sup_{r \in [0,D]} \phi_K(r) < \eta$. This implies

$$\sup_{\theta \in \Theta} \max_{y \in \mathcal{Y}} \{w(\theta, y) - w_K(\theta, y)\} \leq \sup_{r \in [0,D]} \phi_k(r) < \eta.$$

Note that it always holds $w_k(\theta, y) \leq w(\theta, y)$ for each k. Thus we conclude the proof. □

Lemma A.6. *For each n and each closed $G \subset \mathcal{M}$, the following equivalence of the events holds:*

$$\left\{\mu_n^B(z) \in G\right\} = \left\{\lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0 \cap G} \mathcal{Q}_n(\mu, \varepsilon; z) = \lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon; z)\right\}.$$

Proof. ($\subset$). Take any $z \in \mathcal{Z}^n$ such that $\left\{\mu_n^B(z) \in G\right\}$. Note that

$$\lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0 \cap G} \mathcal{Q}_n(\mu, \varepsilon; z) \leq \lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon; z)$$

is clear. Suppose by way of contradiction that

$$\lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0 \cap G} \mathcal{Q}_n(\mu, \varepsilon; z) < \lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon; z).$$

Then, there is small enough $\varepsilon_1 > 0$ such that

$$\max_{\mu \in A_0 \cap G} \mathcal{Q}_n(\mu, \varepsilon_1; z) < \max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon_1; z).$$

We can take $\eta > 0$ such that

$$\max_{\mu \in A_0 \cap G} \mathcal{Q}_n(\mu, \varepsilon_1; z) + 2\eta < \max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon_1; z)$$

Since $\max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon; z) \to \max_{\mu \in A_0} \mathcal{Q}_n(\mu, 0; z)$ as $\varepsilon \downarrow 0$ from Lemmas D.5 and D.6, there exists small enough $\varepsilon_2 > 0$ such that

$$\left|\max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon_2; z) - \max_{\mu \in A_0} \mathcal{Q}_n(\mu, 0; z)\right| < \eta.$$

Also, because $\varepsilon H(\mu_n^B) \to 0$ as $\varepsilon \downarrow 0$, there exists small enough $\varepsilon_3 > 0$ such that $\varepsilon_3 H(\mu_n^B) < \eta$. Let $\varepsilon = \min\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$. Then,

$$\begin{aligned}
\max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon; z) &< \max_{\mu \in A_0} \mathcal{Q}_n(\mu, 0; z) + \eta \\
&= \int\int \sqrt{n}\,(w(\theta_{nh}, x, t) - w(\theta_0, x, t))\, \mathrm{d}\mu_n^B(z) \mathrm{d}\pi(\theta_{nh}|z) + \eta \\
&\leq \int\int \sqrt{n}\,(w(\theta_{nh}, x, t) - w(\theta_0, x, t))\, \mathrm{d}\mu_n^B(z) \mathrm{d}\pi(\theta_{nh}|z) - \varepsilon H(\mu_n^B) + \varepsilon H(\mu_n^B) + \eta \\
&\leq \max_{\mu \in A_0 \cap G} \mathcal{Q}_n(\mu, \varepsilon; z) + \varepsilon H(\mu_n^B) + \eta \\
&< \max_{\mu \in A_0 \cap G} \mathcal{Q}_n(\mu, \varepsilon; z) + 2\eta < \max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon; z),
\end{aligned}$$

which is a contradiction. Hence it holds $\lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0 \cap G} \mathcal{Q}_n(\mu, \varepsilon; z) = \lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon; z)$.

($\supset$). For the other direction, take any $z \in \mathcal{Z}^n$ with

$$\lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0 \cap G} \mathcal{Q}_n(\mu, \varepsilon; z) = \lim_{\varepsilon \downarrow 0} \max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon; z).$$

Let $\varepsilon_k \downarrow 0$ as $k \to \infty$. Recall

$$\mu_{n,\varepsilon_k}^B(z) = \arg\max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon_k; z).$$

By Proposition 3.2, $\mu_{n,\varepsilon_k}^B(z)$ converges to $\mu_n^B(z)$ weakly as $k \to \infty$. Moreover, by Villani (2009, Theorem 6.9), we know that weak convergence in $\mathcal{M}$ is equivalent to convergence in $(\mathcal{M}, d_W)$. Hence, $\mu_{n,\varepsilon_k}^B(z) \to \mu_n^B(z)$ in the Wasserstein distance d_W.

First, we argue that for each $K \in \mathbb{N}$ there is $k \geq K$ such that $\mu_{n,\varepsilon_k}^B(z) \in G$. By way of contradiction, assume that there is K such that for any $k \geq K$, $\mu_{n,\varepsilon_k}^B(z) \notin G$. Now, $\mu_{n,\varepsilon_k}^B(z) \notin G$ implies that

$$\max_{\mu \in A_0 \cap G} \mathcal{Q}_n(\mu, \varepsilon_k; z) < \mathcal{Q}_n(\mu_{n,\varepsilon_k}^B, \varepsilon_k; z).$$

Take any $\eta > 0$ such that

$$\max_{\mu \in A_0 \cap G} \mathcal{Q}_n(\mu, \varepsilon_k; z) + \eta < \mathcal{Q}_n(\mu_{n,\varepsilon_k}^B, \varepsilon_k; z).$$

Because

$$\lim_{k \to \infty} \left(\max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon_k; z) - \max_{\mu \in (A_0 \cap G)} \mathcal{Q}_n(\mu, \varepsilon_k; z) \right) = 0,$$

there exists K' such that if $k \geq K'$ then

$$\left| \max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon_k; z) - \max_{\mu \in (A_0 \cap G)} \mathcal{Q}_n(\mu, \varepsilon_k; z) \right| = \max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon_k; z) - \max_{\mu \in (A_0 \cap G)} \mathcal{Q}_n(\mu, \varepsilon_k; z) < \eta/2.$$

Therefore, for sufficiently large k, we have

$$\max_{\mu \in A_0 \cap G} \mathcal{Q}_n(\mu, \varepsilon_k; z) + \eta < \mathcal{Q}_n(\mu_{n,\varepsilon_k}^B, \varepsilon_k; z) \leq \max_{\mu \in A_0} \mathcal{Q}_n(\mu, \varepsilon_k; z) < \max_{\mu \in (A_0 \cap G)} \mathcal{Q}_n(\mu, \varepsilon_k; z) - \eta/2,$$

which leads to a contradiction. Therefore, for each K there is $k \geq K$ such that $\mu_{n,\varepsilon_k}^B(z) \in G$.

Now, create such a subsequence $\left\{\mu^B_{n,\varepsilon_{k_\ell}}(z)\right\}_{\ell\in\mathbb{N}}$ with $\mu^B_{n,\varepsilon_{k_\ell}}(z) \in G$ for each ℓ. Note that any subsequence of convergent sequence in arbitrary metric space converges to the same limit as the original sequence. Therefore, $\mu^B_{n,\varepsilon_{k_\ell}}(z) \to \mu^B_n(z)$ in d_W as $\ell \to \infty$. Since G is closed, we conclude that $\mu^B_n(z) \in G$. □

Let

$$\begin{aligned} Q_n(\mu, \varepsilon; z) &= \int \left(\int \dot{w}_{\theta_0}(x,t;h)\mathrm{d}\mu\right) \pi_n(\theta_{nh}|z)\mathrm{d}h - \varepsilon H(\mu), \\ \tilde{Q}_n(\mu, \varepsilon; z) &= \int \left(\int \dot{w}_{\theta_0}(x,t;h)\mathrm{d}\mu\right) \mathrm{d}N(\Delta_n(z), I_0^{-1})(h) - \varepsilon H(\mu), \end{aligned}$$

where $\Delta_n(z) = \frac{1}{\sqrt{n}}\sum_{i=1}^n I_0^{-1}s(z_i) \in \mathbb{R}^k$ with the score function s at θ_0 and $\Delta_n \overset{0}{\rightsquigarrow} G \sim N(0, I_0^{-1})$. Define $\mathbb{D} \equiv \mathcal{M} \times [0,1]$. Let $\{Q_n(\mu,\varepsilon) : (\mu,\varepsilon) \in \mathbb{D}\}$, $\{\tilde{Q}_n(\mu,\varepsilon) : (\mu,\varepsilon) \in \mathbb{D}\}$, $\{Q_\infty(\mu,\varepsilon) : (\mu,\varepsilon) \in \mathbb{D}\}$ be stochastic processes. Assume that they yield maps $Q_n : \mathcal{Z}^n \to \ell^\infty(\mathbb{D})$, $\tilde{Q}_n : \mathcal{Z}^n \to \ell^\infty(\mathbb{D})$, and $Q_\infty : \mathbb{R}^k \to \ell^\infty(\mathbb{D})$. We can do this since the sample paths are continuous by Lemma B.2.

Let

$$\mathcal{F} = \left\{ f \in \ell^\infty(\mathbb{D}) : \lim_{\varepsilon\downarrow 0} \max_{\mu\in A_0} f(\mu,\varepsilon) \text{ exists} \right\},$$

where the sup-norm is equipped to $\mathcal{F}$. Note that $Q_n(z), \tilde{Q}_n(z), Q_\infty(\Delta) \in \mathcal{F}$ for any z and Δ.

Lemma A.7. *$\mathcal{Q}_n \overset{h}{\rightsquigarrow} Q_\infty$ in $\mathcal{F}$ as $n \to \infty$.*

Proof. Note that $\mathcal{Q}_n = Q_n + o_{P^n_{\theta_{nh}}}(1)$ as a process in $\mathcal{F}$ as $n \to \infty$ by Lemma A.8. Hence, it is sufficient to show that $Q_n \overset{h}{\rightsquigarrow} Q_\infty$ in $\mathcal{F}$ as $n \to \infty$. Let $C_M \subset \mathbb{R}^k$ be the closed ball of radius M around 0. Define stochastic processes by

$$\begin{aligned} Q_{n,M}(\mu,\varepsilon;z) &= \int_{C_M} \left(\int \dot{w}_{\theta_0}(x,t;h)\mathrm{d}\mu(x,t)\right) \pi_n(\theta_{nh}|z)\mathrm{d}h - \varepsilon H(\mu), \\ \tilde{Q}_{n,M}(\mu,\varepsilon;z) &= \int_{C_M} \left(\int \dot{w}_{\theta_0}(x,t;h)\mathrm{d}\mu(x,t)\right) \mathrm{d}N(\Delta_n(z), I_0^{-1})(h) - \varepsilon H(\mu), \\ Q_{\infty,M}(\mu,\varepsilon;\Delta) &= \int_{C_M} \left(\int \dot{w}_{\theta_0}(x,t;h)\mathrm{d}\mu(x,t)\right) \mathrm{d}N(\Delta, I_0^{-1})(h) - \varepsilon H(\mu). \end{aligned}$$

First, we will show that $Q_{n,M} - \tilde{Q}_{n,M} \overset{h}{\to} 0$ in $\mathcal{F}$ as $n \to \infty$ for any fixed M. Note that

$$\begin{aligned} &\left\| Q_{n,M}(\cdot; Z^n) - \tilde{Q}_{n,M}(\cdot, Z^n) \right\|_{\mathcal{F}} \\ &\le \max_{h\in C_M} \max_{(x,t)\in\mathcal{X}\times\mathcal{T}} |\dot{w}_{\theta_0}(x,t;h)| \cdot \left\| \pi_n(\theta_{nh}|Z^n) - N(\Delta_n(Z^n), I_0^{-1}) \right\|_{\mathrm{TV}} \end{aligned}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm. Because C_M is bounded and $\dot{w}_{\theta_0}(x,t;h)$ is bounded for any h, we know that $\max_{h\in C_M} \max_{(x,t)\in\mathcal{X}\times\mathcal{T}} |\dot{w}_{\theta_0}(x,t;h)| < \infty$. From the Bernstein–von Mises theorem,

$$\left\| \pi_n(\theta_{nh}|Z^n) - N(\Delta_n(Z^n), I_0^{-1}) \right\|_{\mathrm{TV}} \overset{0}{\to} 0 \quad \text{as } n \to \infty.$$

From Le Cam's first lemma, it also converges to 0 along $P^n_{\theta_{nh}}$. Therefore, $Q_{n,M} - \tilde{Q}_{n,M} \xrightarrow{h} 0$ in $\mathcal{F}$ as $n \to \infty$.

Next, we argue that $\tilde{Q}_{n,M} \overset{h}{\rightsquigarrow} Q_M$ in $\mathcal{F}$ as $n \to \infty$ for any fixed M. Define $\phi : \mathbb{R}^k \to \mathcal{F}$ by

$$\phi(\delta)(\mu, \varepsilon) = \int_{C_M} \left(\int \dot{w}_{\theta_0}(x,t;h) \mathrm{d}\mu(x,t) \right) \mathrm{d}N(\delta, I_0^{-1})(h) - \varepsilon H(\mu)$$

Since ϕ is continuous, and $\Delta_n \overset{h}{\rightsquigarrow} \Delta \sim N(h, I_0^{-1})$ as $n \to \infty$ by Le Cam's third lemma, the continuous mapping theorem implies $\phi(\Delta_n) \overset{h}{\rightsquigarrow} \phi(\Delta)$ as $n \to \infty$. Thus $\tilde{Q}_{n,M} \overset{h}{\rightsquigarrow} Q_M$ in $\mathcal{F}$ as $n \to \infty$.

Combining the above two findings, we obtain $Q_{n,M} \overset{h}{\rightsquigarrow} Q_{\infty,M}$ in $\mathcal{F}$ as $n \to \infty$ from the Slutsky theorem. We also have that $Q_{\infty,M} - Q_\infty = o_{P_h^\Delta}(1)$ as $M \to \infty$ where P_h^Δ is the (marginal) law of $\Delta \sim N(h, I_0^{-1})$. Thus, there exists a sequence $M_n \to \infty$ such that $Q_{n,M_n} \overset{h}{\rightsquigarrow} Q_\infty$ in $\mathcal{F}$ as $n \to \infty$.

Finally, it remains to show that $Q_n - Q_{n,M_n} = o_{P^n_{\theta_{nh}}}(1)$ in $\mathcal{F}$ as $n \to \infty$, which leads to the conclusion, $Q_n \overset{h}{\rightsquigarrow} Q_\infty$ in $\mathcal{F}$ as $n \to \infty$. By Assumption 3.4 (iii),

$$\left| \int_{\mathbb{R}^k \setminus C_{M_n}} \left(\int \dot{w}_{\theta_0}(x,t;h) \mathrm{d}\mu(x,t) \right) \pi_n(\theta_{nh}|z) \mathrm{d}h \right| \leq \int_{\mathbb{R}^k \setminus C_{M_n}} K(h) \pi_n(\theta_{nh}|z) \mathrm{d}h.$$

Then applying Proposition A.1 yields that RHS is $o_{P^n_{\theta_0}}(1)$. Thus it is $o_{P^n_{\theta_{nh}}}(1)$ as well. Hence it follows that $Q_n(\mu, \varepsilon; Z^n) - Q_{n,M_n}(\mu, \varepsilon; Z^n) \xrightarrow{h} 0$ for any $(\mu, \varepsilon) \in \mathbb{D}$. Then the continuity of sample path implies $Q_n - Q_{n,M_n} = o_{P^n_{\theta_{nh}}}(1)$ in $\mathcal{F}$ as desired. □

Lemma A.8. *$\mathcal{Q}_n = Q_n + o_{P^n_{\theta_{nh}}}(1)$ in $\mathcal{F}$ as $n \to \infty$.*

Proof. Notice that for every $z \in \mathcal{Z}^n$, $\sup_{(\mu,\varepsilon) \in \mathbb{D}} |\mathcal{Q}_n(\mu, \varepsilon; z) - Q_n(\mu, \varepsilon; z)|$ is bounded above by

$$\int \max_{(x,t) \in \mathcal{X} \times \mathcal{T}} |\tilde{w}_{n,0}(x,t;h)| \, \pi_n(\theta_{nh}|z) \mathrm{d}h,$$

where $\tilde{w}_{n,0}(x,t;h) := \sqrt{n}\,(w(\theta_{nh}, x, t) - w(\theta_0, x, t)) - \dot{w}_{\theta_0}(x,t;h)$. Let C_{M_n} be a closed ball of radius M_n around 0, where M_n is the divergent sequence specified in Lemma B.4. Then the previous display is further bounded by

$$\pi_n(\theta_{nh}|z)(C_{M_n}) \times \max_{h \in C_{M_n}} \max_{(x,t) \in \mathcal{X} \times \mathcal{T}} |\tilde{w}_{n,0}(x,t;h)| + \int \mathbf{1}_{\mathbb{R}^k \setminus C_{M_n}}(h) \max_{(x,t) \in \mathcal{X} \times \mathcal{T}} |\tilde{w}_{n,0}(x,t;h)| \, \pi_n(\theta_{nh}|z) \mathrm{d}h \tag{A.8}$$

The first term of (A.8) converges to zero as $n \to \infty$ by Lemma B.4. For the second term of (A.8), note that from Assumption 3.4 (i),

$$\max_{(x,t) \in \mathcal{X} \times \mathcal{T}} \left| \sqrt{n}\,(w(\theta_{nh}, x, t) - w(\theta_0, x, t)) \right| \leq \max_{(x,t) \in \mathcal{X} \times \mathcal{T}} |\dot{w}_{\theta_0}(x,t;h)| + o(1),$$

From Assumption 3.4 (iii), $\max_{(x,t) \in \mathcal{X} \times \mathcal{T}} |\dot{w}_{\theta_0}(x,t;h)|$ is bounded by $K(h)$ that grows at subpolynomially of order p. This implies that $\max_{(x,t) \in \mathcal{X} \times \mathcal{T}} |\tilde{w}_{n,0}(x,t;h)|$ is also dominated by a function that grows subpolynomially of order p for sufficiently large n. Then applying Proposition A.1 yields the conclusion. □

Appendix B. Auxiliary lemmas for Theorem 3.3

The following extends the result on Hadamard directional differentiability given by Römisch (2004, Proposition 1). Compared to his setting, the objective map $\int w(\theta, x, t)\mathrm{d}\mu$ need not to be linear in θ. By leveraging the uniform continuity from Assumptions 3.3 and 3.4, we obtain the similar form of the directional derivative as his result.

Lemma B.1. *For a closed set $S \subset \mathcal{M}$, define the map $W_S^* : \Theta \to \mathbb{R}$ by*

$$W_S^*(\theta) = \max_{\mu \in S} \int w(\theta, x, t)\mathrm{d}\mu.$$

Then W_S^* is Hadamard directionally differentiable with derivative

$$\dot{W}_{S,0}^*[h] \equiv \lim_{\varepsilon \downarrow 0} \sup_{\mu \in S^\varepsilon(\theta_0)} \int \dot{w}_{\theta_0}(x, t; h)\mathrm{d}\mu$$

where

$$S^\varepsilon(\theta) \equiv \left\{\mu \in S : \int w(\theta, x, t)\mathrm{d}\mu + \varepsilon \geq \max_{\mu \in S} \int w(\theta, x, t)\mathrm{d}\mu\right\} \neq \emptyset$$

for $\varepsilon > 0$ and $\theta \in \Theta$. Moreover, if $S \subset A_0$ then $\dot{W}_{S,0}^*[h] = \max_{\mu \in S} \int \dot{w}_{\theta_0}(x, t; h)\mathrm{d}\mu$.

Proof. The second statement follows from the first statement and the fact that $S^\varepsilon(\theta_0) = S$ for any $\varepsilon > 0$. Hereafter, we focus on the first statement.

Fix any closed $S \subset \mathcal{M}$ and any $\theta_0 \in \Theta$. Then, $\arg\max_{\mu \in S} W(\theta_0, \mu) \neq \emptyset$ because $w(\theta_0, \cdot)$ is bounded continuous on $\mathcal{X} \times \mathcal{T}$ from Assumption 3.3 (Villani, 2009, Theorem 4.1). Let $\mu(\theta_0) \in \arg\max_{\mu \in S} W(\theta_0, \mu)$. Since $\arg\max_{\mu \in S} W(\theta_0, \mu) \subset S^\varepsilon(\theta_0)$ for any $\varepsilon > 0$, we can guarantee $S^\varepsilon(\theta_0) \neq \emptyset$. Also, because $w(\theta_0, \cdot, \cdot)$ is bounded, $\max_{\mu \in S} \int w(\theta_0, x, t)\mathrm{d}\mu < \infty$.

Let $r_n \downarrow 0$, and $h_n \to h$. Define

$$\sigma_n = \frac{1}{r_n}\left(W_S^*(\theta_0 + r_n h_n) - W_S^*(\theta_0)\right)$$

and we want to show that $\sigma_n \to \dot{W}_{S,0}^*[h]$. But, it is enough to show that for any subsequence of $\{\sigma_n\}$, there exists a further subsequence that converges to $\dot{W}_{S,0}^*[h]$. Take any subsequence and denote it by $\{\sigma_n\}$ for simplicity.

Define the maps $T_n : S \to \mathbb{R}$ and $T : S \to \mathbb{R}$ by

$$T_n(\mu) = \int \frac{1}{r_n}\left(w(\theta_0 + r_n h_n, x, t) - w(\theta_0, x, t)\right)\mathrm{d}\mu,$$
$$T(\mu) = \int \dot{w}_{\theta_0}(x, t; h)\mathrm{d}\mu.$$

First, to see that $T_n \to T$ uniformly, take any $\mu \in S$. Note that

$$|T_n(\mu) - T(\mu)| \leq \max_{(x,t) \in \mathcal{X} \times \mathcal{T}} \left|\frac{1}{r_n}\left(w(\theta_0 + r_n h_n, x, t) - w(\theta_0, x, t)\right) - \dot{w}_{\theta_0}(x, t; h)\right|.$$

We can make RHS arbitrary small as $n \to \infty$ without depending on μ from Assumption 3.4 (i). Hence, we conclude that $T_n \to T$ uniformly.

Thus, for each n there exists K_1 such that for each $k \geq K_1$

$$\left| \int \frac{1}{r_{n_k}} \left(w(\theta_0 + r_{n_k} h_{n_k}, x, t) - w(\theta_0, x, t) \mathrm{d}\mu \right) - \int \dot{w}_{\theta_0}(x, t; h) \mathrm{d}\mu \right| < r_n^2 \quad \forall \mu \in S. \tag{B.1}$$

Moreover, from Assumption 3.3 (i), there exists K_2 such that for each $k \geq K_2$

$$\left| \int w(\theta_0 + r_{n_k} h_{n_k}, x, t) \mathrm{d}\mu - \int w(\theta_0, x, t) \mathrm{d}\mu \right| < r_n^2/2 \quad \forall \mu \in S. \tag{B.2}$$

Let $k \geq \max\{K_1, K_2\}$, and construct a further subsequence $\{\sigma_{n_k}\}$.

For each k, take any $\mu_{n_k} \in S^{r_n^2 r_{n_k}}(\theta_0)$. Then, from $\mu_{n_k} \in S$ and the definition of $S^{r_n^2 r_{n_k}}(\theta_0)$,

$$\begin{aligned} \frac{1}{r_{n_k}} \left(W_S^*(\theta_0 + r_{n_k} h_{n_k}) - W_S^*(\theta_0) \right) &= \frac{1}{r_{n_k}} \left(\max_{\mu \in S} \int w(\theta_0 + r_{n_k} h_{n_k}, x, t) \mathrm{d}\mu - \max_{\mu \in S} \int w(\theta_0, x, t) \mathrm{d}\mu \right) \\ &\geq \frac{1}{r_{n_k}} \left(\int w(\theta_0 + r_{n_k} h_{n_k}, x, t) \mathrm{d}\mu_{n_k} - \max_{\mu \in S} \int w(\theta_0, x, t) \mathrm{d}\mu \right) \\ &\geq \frac{1}{r_{n_k}} \left(\int w(\theta_0 + r_{n_k} h_{n_k}, x, t) \mathrm{d}\mu_{n_k} - \int w(\theta_0, x, t) \mathrm{d}\mu_{n_k} \right) - r_n^2. \end{aligned}$$

From (B.1), we have

$$\sigma_{n_k} > \int \dot{w}_0(x, t; h) \mathrm{d}\mu_{n_k} - 2r_n^2.$$

Therefore,

$$\sigma_{n_k} \geq \sup_{\mu \in S^{r_n^2 r_{n_k}}(\theta_0)} \int \dot{w}_0(x, t; h) \mathrm{d}\mu - 2r_n^2.$$

Since $r_n^2 \downarrow 0$, it leads to

$$\liminf_{k \to \infty} \sigma_{n_k} \geq \lim_{\varepsilon \downarrow 0} \sup_{\mu \in S^{\varepsilon}(\theta_0)} \int \dot{w}_0(x, t; h) \mathrm{d}\mu.$$

Also, take any $\mu'_{n_k} \in S^{r_n^2 r_{n_k}}(\theta_0 + r_{n_k} h_{n_k})$. Then,

$$\begin{aligned} \frac{1}{r_{n_k}} \left(W_S^*(\theta_0 + r_{n_k} h_{n_k}) - W_S^*(\theta_0) \right) &= \frac{1}{r_{n_k}} \left(\max_{\mu \in S} \int w(\theta_0 + r_{n_k} h_{n_k}, x, t) \mathrm{d}\mu - \max_{\mu \in S} \int w(\theta_0, x, t) \mathrm{d}\mu \right) \\ &\leq \frac{1}{r_{n_k}} \left(\int w(\theta_0 + r_{n_k} h_{n_k}, x, t) \mathrm{d}\mu_{n_k} - \int w(\theta_0, x, t) \mathrm{d}\mu_{n_k} \right) + r_n^2. \end{aligned}$$

From (B.1),

$$\sigma_{n_k} < \int \dot{w}_{\theta_0}(x, t; h) \mathrm{d}\mu'_{n_k} + 2r_n^2.$$

If we would have $S^{r_n^2 r_{n_k}}(\theta_0 + r_{n_k} h_{n_k}) \subset S^{r_n^2 r_{n_k} + r_n^2}(\theta_0)$, then we obtain

$$\sigma_{n_k} \leq \sup \left\{ \int \dot{w}_{\theta_0}(x, t; h) \mathrm{d}\mu : \mu \in S^{r_n^2 r_{n_k} + r_n^2}(\theta_0) \right\} + 2r_n^2,$$

which leads to

$$\limsup_{k \to \infty} \sigma_{n_k} \leq \lim_{\varepsilon \downarrow 0} \sup_{\mu \in S^{\varepsilon}(\theta_0)} \int \dot{w}_{\theta_0}(x, t; h) \mathrm{d}\mu,$$

thus $\sigma_{n_k} \to \dot{W}^*_{S,0}[h]$. Hence, it suffices to show $S^{r_n^2 r_{n_k}}(\theta_0+r_{n_k}h_{n_k}) \subset S^{r_n^2 r_{n_k}+r_n^2}(\theta_0)$ for the conclusion. Take any $\nu \in S^{r_n^2 r_{n_k}}(\theta_0 + r_{n_k}h_{n_k})$, then

$$\int w(\theta_0 + r_{n_k}h_{n_k})\mathrm{d}\nu + r_n^2 r_{n_k} \geq \max_{\mu\in S} \int w(\theta_0 + r_{n_k}h_{n_k})\mathrm{d}\mu. \tag{B.3}$$

Let $\mu(\theta_0) \in \arg\max_{\mu\in S} \int w(\theta_0, x, t)\mathrm{d}\mu$, then

$$\begin{aligned}
&\int w(\theta_0, x, t)\mathrm{d}\nu + r_n^2 r_{n_k} + r_n^2 \\
\geq &\int w(\theta_0, x, t)\mathrm{d}\nu + r_n^2 + \max_{\mu\in S}\int w(\theta_0 + r_{n_k}h_{n_k}, x, t)\mathrm{d}\mu - \int w(\theta_0 + r_{n_k}h_{n_k}, x, t)\mathrm{d}\nu \quad \because (\text{B.3}) \\
\geq &\int w(\theta_0, x, t)\mathrm{d}\nu + r_n^2 + \int w(\theta_0 + r_{n_k}h_{n_k}, x, t)\mathrm{d}\mu(\theta_0) - \int w(\theta_0 + r_{n_k}h_{n_k}, x, t)\mathrm{d}\nu \\
= &\int w(\theta_0, x, t)\mathrm{d}\nu + r_n^2 \\
&+ \int w(\theta_0 + r_{n_k}h_{n_k}, x, t)\mathrm{d}\mu(\theta_0) - \int w(\theta_0, x, t)\mathrm{d}\mu(\theta_0) + \int w(\theta_0, x, t)\mathrm{d}\mu(\theta_0) - \int w(\theta_0 + r_{n_k}h_{n_k}, x, t)\mathrm{d}\nu \\
= &\int w(\theta_0, x, t)\mathrm{d}\mu(\theta_0) + r_n^2 \\
&+ \int (w(\theta_0, x, t) - w(\theta_0 + r_{n_k}h_{n_k}, x, t))\,\mathrm{d}\nu + \int (w(\theta_0 + r_{n_k}h_{n_k}, x, t) - w(\theta_0, x, t))\,\mathrm{d}\mu(\theta_0) \\
\geq &\int w(\theta_0, x, t)\mathrm{d}\mu(\theta_0) + r_n^2 - r_n^2/2 - r_n^2/2 \quad \because (\text{B.2}) \\
= &\int w(\theta_0, x, t)\mathrm{d}\mu(\theta_0).
\end{aligned}$$

Thus, $\nu \in S^{r_n^2 r_{n_k}+r_n^2}(\theta_0)$. $\square$

Lemma B.2. *The sample paths of Q_n, $\tilde{Q}_n$, and Q_∞ are continuous and bounded in $\mathcal{M}$.*

Proof. First, we will show that the sample path of

$$Q_n(\mu, \varepsilon; z) = \int \left(\int \dot{w}_{\theta_0}(x, t; h)\mathrm{d}\mu\right) \pi_n(\theta_{nh}|z)\mathrm{d}h - \varepsilon H(\mu)$$

is continuous. Note that, then, it is bounded because $\mathcal{M}$ is compact. Fix any z, and take any $\{(\mu_k, \varepsilon_k)\}_{k=1}^\infty \subset \mathbb{D}$ that converges to (μ, ε). Since $\mathbb{D}$ is a metric space, overall convergence implies elementwise convergence. Thus, $\mu_k \to \mu$ in the Wasserstein distance and $\varepsilon_k \to \varepsilon$. We are done if

$$\left|\int \left\{\int \dot{w}_{\theta_0}(x, t; h)\mathrm{d}\mu_k - \int \dot{w}_{\theta_0}(x, t; h)\mathrm{d}\mu\right\} \pi_n(\theta_{nh}|z)\mathrm{d}h - \varepsilon_k H(\mu_k) + \varepsilon H(\mu)\right| \to 0,$$

as $k \to \infty$.

By the triangle inequality, LHS is bounded above by

$$\left|\int \left\{\int \dot{w}_{\theta_0}(x, t; h)\mathrm{d}\mu_k - \int \dot{w}_{\theta_0}(x, t; h)\mathrm{d}\mu\right\} \pi_n(\theta_{nh}|z)\mathrm{d}h\right| + \left|\varepsilon_k H(\mu_k) - \varepsilon H(\mu)\right|.$$

The second term converges to zero since H is continuous. For the first term, since $\mu_k \to \mu$ in the Wasserstein distance implies $\mu_k \rightsquigarrow \mu$, we have

$$\int \left(\int \dot{w}_{\theta_0}(x,t;h) \mathrm{d}\mu_k \right) \pi_n(\theta_{nh}|z) \mathrm{d}h = \int \left(\int \dot{w}_{\theta_0}(x,t;h) \pi_n(\theta_{nh}|z) \mathrm{d}h \right) \mathrm{d}\mu_k$$
$$\to \int \left(\int \dot{w}_{\theta_0}(x,t;h) \pi_n(\theta_{nh}|z) \mathrm{d}h \right) \mathrm{d}\mu,$$

where the convergence holds if the map $(x,t) \mapsto \int \dot{w}_{\theta_0}(x,t;h)\pi_n(\theta_{nh}|z)\mathrm{d}h$ is continuous and bounded. To see the continuity, let $\{(x_k,t_k)\}$ be such that $(x_k,t_k) \to (x,t)$. Then

$$\left| \int \dot{w}_{\theta_0}(x,t;h)\pi_n(\theta_{nh}|z)\mathrm{d}h - \int \dot{w}_{\theta_0}(x,t;h)\pi_n(\theta_{nh}|z)\mathrm{d}h \right|$$
$$\leq \int |\dot{w}_{\theta_0}(x,t;h) - \dot{w}_{\theta_0}(x,t;h)| \, \pi_n(\theta_{nh}|z)\mathrm{d}h \to 0 \quad \text{as } k \to \infty,$$

where the convergence follows from the continuity of $\dot{w}_{\theta_0}(x,t;h)$ in (x,t) and the dominated convergence theorem. The boundedness of $(x,t) \mapsto \int \dot{w}_{\theta_0}(x,t;h)\pi_n(\theta_{nh}|z)\mathrm{d}h$ follows from the continuity of $\dot{w}_{\theta_0}(x,t;h)$ in (x,t) (Assumption 3.4 (ii)) and the compactness of $\mathcal{X} \times \mathcal{T}$ (Assumption 3.2).

Similar arguments can be applied to $\tilde{Q}_n$ and Q_∞. □

Lemma B.3. *For any $S \subset A_0$. the operator $M : \mathcal{F} \to \mathbb{R}$ where $M(f) := \lim_{\varepsilon \downarrow 0} \max_{\mu \in S} f(\mu, \varepsilon)$ is continuous.*

Proof. Let $f_k \to f$ in $\mathcal{F}$ as $k \to \infty$. Note that $M(f)$ and $M(f_k)$ exist for each k by the definition of $\mathcal{F}$. Then

$$|M(f) - M(f_k)| = \lim_{\varepsilon \downarrow 0} \left| \max_{\mu \in \mathcal{M}} f(\mu,\varepsilon) - \max_{\mu \in \mathcal{M}} f_k(\mu,\varepsilon) \right|$$
$$\leq \lim_{\varepsilon \downarrow 0} \max_{\mu \in \mathcal{M}} |f(\mu,\varepsilon) - f_k(\mu,\varepsilon)|$$
$$\leq \max_{(\mu,\varepsilon) \in \mathbb{D}} |f(\mu,\varepsilon) - f_k(\mu,\varepsilon)| \to 0,$$

as $k \to \infty$, where the equality follows because $|\cdot|$ is continuous, and the convergence follows because $f_k \to f$ in $\mathcal{F}$ as $k \to \infty$. □

The next result is an adaptation of Christensen *et al.* (2025, Lemma 3), where we need a modification to allow max-operator within the expression.

Lemma B.4. *There is a sequence $\{M_n\}$ such that $M_n \uparrow \infty$, $M_n/\sqrt{n} \to 0$, and*

$$\sup_{\|h\| \leq 2M_n} \max_{(x,t) \in \mathcal{X} \times \mathcal{T}} \left| \sqrt{n} \left[w(\theta_{nh},x,t) - w(\theta_0,x,t) \right] - \dot{w}_{\theta_0}(x,t;h) \right| \to 0.$$

Proof. From Shapiro (1990, Lemmas 3.3 and 3.4) and Assumption 3.4 (i), we know that for any compact $S \subset \mathbb{R}^k$ and $\varepsilon > 0$, there is N such that $\sup_{h \in S} g_n(h) < \varepsilon$ for any $n \geq N$ where $g_n(h) = \max_{(x,t) \in \mathcal{X} \times \mathcal{T}} |\sqrt{n} \left[w(\theta_{nh},x,t) - w(\theta_0,x,t) \right] - \dot{w}_{\theta_0}(x,t;h)|$. Define

$$\psi_n = \sup_{\|h\| \leq 2\log(1+n)} g_n(h).$$

We are done if $\psi_n \to 0$ because $\log(1+n) \uparrow \infty$ and $n^{-1/2}\log(n+1) \to 0$. To show $\psi_n \to 0$, it is enough to show that for any subsequence ψ_n (abusing notation) there is a further subsequence converging to 0. First, consider $\sup_{\|h\|\le 2\log(1+1)} g_n(h)$. We know that there is $N(1)$ such that $\sup_{\|h\|\le 2\log(1+1)} g_n(h) < 1/\log(1+1)$ for any $n \ge N(1)$. Second, for $\sup_{\|h\|\le 2\log(1+2)} g_n(h)$, there is $N(2)$ such that $\sup_{\|h\|\le 2\log(1+2)} g_n(h) < 1/\log(1+2)$ for any $n \ge N(2)$. Proceed with $N(1) < N(2) < N(3) < \cdots$ WLOG. Then, the map $N : \mathbb{N} \to \mathbb{N}$ satisfies $i < j \implies N(i) < N(j)$. Hence, $\psi_{N(n)}$ is a subsequence of ψ_n. To show $\psi_{N(n)} = \sup_{\|h\|\le 2\log(1+n)} g_{N(n)}(h) \to 0$, take any $\varepsilon > 0$. Since $1/\log(1+n) \downarrow 0$, there is $\bar{n}$ such that $1/\log(1+n) < \varepsilon$ for any $n \ge \bar{n}$. Therefore, $\psi_{N(n)} < 1/\log(1+n) < \varepsilon$ for any $n \ge \bar{n}$. □

Appendix C. Proof of Proposition 3.1

By Villani (2009, Corollary 6.11), $\mu \mapsto d_W(\mu,\nu)$ is continuous on $\mathcal{M}$.[13] Convexity of $\mu \mapsto d_W(\mu,\nu)$ is shown below. Because the map $\mu \mapsto d_W(\mu,\nu)$ is convex and nonnegative and the map $\mathbb{R}_+ \ni x \mapsto x^2$ is increasing and strictly convex, the composite map $\mu \mapsto (d_W(\mu,\nu))^2$ is strictly convex and nonnegative. Also, it is bounded by the compactness of $\mathcal{M}$ and the continuity.

To see the convexity of $\mu \mapsto d_W(\mu,\nu)$, fix any $\mu_1,\mu_2 \in \mathcal{M}$ and $\alpha \in (0,1)$. To simplify the notation, let $\mathcal{Y} = \mathcal{X} \times \mathcal{T}$. Let

$$\gamma_1^* \in \arg\inf_{\gamma\in\Gamma(\mu_1,\nu)} \int d(y,y')\gamma(\mathrm{d}y,\mathrm{d}y'),$$
$$\gamma_2^* \in \arg\inf_{\gamma\in\Gamma(\mu_2,\nu)} \int d(y,y')\gamma(\mathrm{d}y,\mathrm{d}y').$$

First, we show that $\alpha\gamma_1^* + (1-\alpha)\gamma_2^* \in \Gamma(\alpha\mu_1 + (1-\alpha)\mu_2, \nu)$. Take any $A, B \in \mathcal{Y}$. Then,

$$\left(\alpha\gamma_1^* + (1-\alpha)\gamma_2^*\right)(\mathcal{Y}\times B) = \alpha\gamma_1^*(\mathcal{Y}\times B) + (1-\alpha)\gamma_2^*(\mathcal{Y}\times B) = \nu(B).$$

Also,

$$\begin{aligned}&\left(\alpha\gamma_1^* + (1-\alpha)\gamma_2^*\right)(A\times\mathcal{Y})\\ &= \alpha\gamma_1^*(A\times\mathcal{Y}) + (1-\alpha)\gamma_2^*(A\times\mathcal{Y}) = \alpha\mu_1(A) + (1-\alpha)\mu_2(A) = \left(\alpha\mu_1 + (1-\alpha)\mu_2\right)(A).\end{aligned}$$

Hence,

$$\begin{aligned}r(\alpha\mu_1 + (1-\alpha)\mu_2) &\le \int d(y,y')\mathrm{d}(\alpha\gamma_1^* + (1-\alpha)\gamma_2^*)\\ &= \alpha\int d(y,y')\mathrm{d}\gamma_1^* + (1-\alpha)\int d(y,y')\mathrm{d}\gamma_2^* = \alpha r(\mu_1) + (1-\alpha)r(\mu_2).\end{aligned}$$

[13]More explicitly, if μ_k converges to μ weakly in $\mathcal{M}$, then $d_W(\mu_k,\nu) \to d_W(\mu,\nu)$.

Appendix D. Proof of Proposition 3.2

We provide a proof under a general framework using continuous and bounded cost function $c : \mathcal{X} \times \mathcal{T} \to \mathbb{R}$. Define

$$\mathcal{C}_\varepsilon := \inf_{\mu \in \mathcal{M}} \int c \mathrm{d}\mu + \varepsilon H(\mu). \qquad (\varepsilon\text{EOT})$$

$$\mathcal{C}_0 := \inf_{\mu \in \mathcal{M}} \int c \mathrm{d}\mu. \qquad (\text{OT}).$$

Let $\mathcal{M}_{opt} = \arg\min_{\mu \in \mathcal{M}} \int c \mathrm{d}\mu$. It should be noted that $(\mathcal{M}, d_W)$ is a metric space which is convex and compact.

Our proof of Proposition 3.2 proceeds as follows. Lemma D.4 gives the conclusion under the assumption $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$. Lemmas D.1–D.3 are needed to prove Lemma D.4. Finally, Lemmas D.5 and D.6 show $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$.

Lemma D.1. *Let $\mu_n \in \mathcal{M}$. Suppose that $\lim_n H(\mu_n) =: a \in \mathbb{R}$ exists and that*

$$\limsup_{m,n \to \infty} H(\mu_{m,n}) \geq a$$

for $\mu_{m,n} := (\mu_m + \mu_n)/2$. Then $\{\mu_n\}$ converges weakly.

Proof. Let $D < \infty$ be the diameter of $\mathcal{X} \times \mathcal{T}$. By Villani (2009, Theorem 6.15), $d_W(\mu, \nu)$ is bounded by $D \|\mu - \nu\|_{\mathrm{TV}}$. By following the same arguments of Nutz (2022, Lemma 1.9), we obtain

$$\lim_{m,n \to \infty} \|\mu_m - \mu_n\|_{\mathrm{TV}} = 0.$$

Thus it follows that $\lim_{m,n \to \infty} d_W(\mu_m, \mu_n) = 0$. □

The next result is an adaptation of Nutz (2022, Theorem 1.10). We use weak convergence as the mode of convergence, whereas the original proof uses convergence in total variation.

Lemma D.2. *Let $\mathcal{Q} \subset \mathcal{M}$ be a convex and closed subset. There exists a unique $\mu_* \in \mathcal{Q}$ such that*

$$H(\mu_*) = \inf_{\mu \in \mathcal{Q}} H(\mu) \in [0, \infty).$$

Proof. Let $\mu_n \in \mathcal{Q}$ be such that $H(\mu_n) \to \inf_{\mu' \in \mathcal{Q}} H(\mu')$. By convexity of $\mathcal{Q}$, we have $\mu_{m,n} := (\mu_m + \mu_n)/2 \in \mathcal{Q}$ and hence $H(\mu_{m,n}) \geq \inf_{\mu \in \mathcal{Q}} H(\mu)$ for all m, n. Lemma D.1 shows that $\{\mu_n\}$ converges weakly to some μ_*. By the continuity of $\mu \mapsto H(\mu)$, μ_* is a minimizer of $\inf_{\mu' \in \mathcal{Q}} H(\mu')$. Uniqueness follows from the strict convexity of H. □

The next result is an adaptation of Nutz (2022, Proposition 1.17).

Lemma D.3. *Consider a decreasing sequence of sets $\mathcal{Q}_n \subset \mathcal{M}$ that are convex and closed, and let $\mathcal{Q} := \cap_n \mathcal{Q}_n$. Let $\mu_n = \arg\min_{\mu \in \mathcal{Q}_n} H(\mu)$ be the minimizer of $\mathcal{Q}_n$. Then*

$$\mu_n \to \mu_* \text{ weakly}, \quad \text{and} \quad H(\mu_n) \to H(\mu_*),$$

where $\mu_* = \arg\min_{\mu' \in \mathcal{Q}} H(\mu')$.

Proof. Note that the inclusion $\mathcal{Q}_n \supset \mathcal{Q}_{n+1} \supset \mathcal{Q}$ implies that $H(\mu_n)$ is increasing and $H(\mu_n) \leq \inf_{\mu' \in \mathcal{Q}} H(\mu')$. Since any increasing and bounded-above sequence is convergent, we have $\lim H(\mu_n) \leq \inf_{\mu' \in \mathcal{Q}} H(\mu') < \infty$. For $m \geq n$, we have $\mu_{m,n} := (\mu_m + \mu_n)/2 \in \mathcal{Q}_n$ by convexity. Then $H(\mu_{m,n}) \geq H(\mu_n)$. Thus $\limsup_{m,n \to \infty} H(\mu_{m,n}) \geq \lim H(\mu_n)$. Since $\lim H(\mu_n) < \infty$, Lemma D.1 implies that μ_n converges weakly to some limit μ. By the continuity of H on $\mathcal{M}$,

$$H(\mu) = \lim_n H(\mu_n) \leq \inf_{\mu' \in \mathcal{Q}} H(\mu').$$

Thus we obtain $\mu \in \arg\min_{\mu' \in \mathcal{Q}} H(\mu')$. By the uniqueness of the minimizer shown in Lemma D.2, we have $\mu = \mu_*$; i.e., μ_n converges weakly to μ_*. □

The next result is an adaptation of Nutz (2022, Theorem 5.5).

Lemma D.4. *Suppose that* $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$. *Let* μ_ε *be the optimizer of (εEOT). Then,*

$$\mu_\varepsilon \to \mu_* \textit{ weakly as } \varepsilon \downarrow 0, \quad \textit{and} \quad H(\mu_\varepsilon) \to H(\mu_*),$$

where $\mu_* = \arg\min_{\mu \in \mathcal{M}_{opt}} H(\mu)$.

Proof. The additive form of (εEOT) and the optimality of the couplings imply that

$$H(\mu_\varepsilon) \leq H(\mu_{\varepsilon'}) \quad \text{and} \quad \int c \mathrm{d}\mu_\varepsilon \geq \int c \mathrm{d}\mu_{\varepsilon'} \quad \text{for } \varepsilon \geq \varepsilon' > 0.$$

Denote $\mathcal{Q} := \mathcal{M}_{opt}$ and

$$\mathcal{Q}_\varepsilon := \left\{ \mu \in \mathcal{M} : \int c \mathrm{d}\mu \leq \int c \mathrm{d}\mu_\varepsilon \right\}.$$

Note that $\mathcal{Q}_\varepsilon$ is a closed convex set, and $\mu_\varepsilon = \arg\min_{\mu \in \mathcal{Q}_\varepsilon} H(\mu)$.[14] Then $\int c \mathrm{d}\mu_\varepsilon \geq \int c \mathrm{d}\mu_{\varepsilon'}$ implies that $\mathcal{Q}_\varepsilon \supset \mathcal{Q}_{\varepsilon'}$ for $\varepsilon \geq \varepsilon'$. Next, we claim that $\mathcal{Q} = \cap_\varepsilon \mathcal{Q}_\varepsilon$. It is easy to see $\mathcal{Q} \subset \cap_\varepsilon \mathcal{Q}_\varepsilon$. For the other direction, take any $\mu \in \cap_\varepsilon \mathcal{Q}_\varepsilon$. Then we have $\mu \in \mathcal{Q}$ because

$$\int c \mathrm{d}\mu \leq \int c \mathrm{d}\mu_\varepsilon \leq \mathcal{C}_\varepsilon \to \mathcal{C}_0.$$

Then applying Lemma D.3 completes the proof. □

Thus, it remains to show that $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$. The next result is an adaptation of Nutz (2022, Lemma 5.2).

Lemma D.5. *Suppose that given* $\eta > 0$, *there exists* $\mu^\eta \in \mathcal{M}$ *with* $\int c \mathrm{d}\mu^\eta \leq \mathcal{C}_0 + \eta$ *and* $H(\mu^\eta) < \infty$. *Then* $\lim_{\varepsilon \to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$.

Proof. Given $\eta > 0$, we have

$$\mathcal{C}_\varepsilon \leq \int c \mathrm{d}\mu^\eta + \varepsilon H(\mu^\eta) \leq \mathcal{C}_0 + \eta + \varepsilon H(\mu^\eta).$$

[14] Suppose, by contradiction, that there exists $\mu \in \mathcal{Q}_\varepsilon$ such that $H(\mu) < H(\mu_\varepsilon)$. Then

$$\int c \mathrm{d}\mu + \varepsilon H(\mu) < \int c \mathrm{d}\mu_\varepsilon + \varepsilon H(\mu_\varepsilon),$$

which contradicts with the optimality of μ_ε.

Thus $\lim_{\varepsilon\to 0} \mathcal{C}_\varepsilon \le \mathcal{C}_0 + \eta$. Since $\eta > 0$ is arbitrary, we are done. ☐

The next result is an adaptation of Nutz (2022, Lemma 5.4).

Lemma D.6. *Let c be continuous and bounded. Then* $\lim_{\varepsilon\to 0} \mathcal{C}_\varepsilon = \mathcal{C}_0$.

Proof. Let $\eta > 0$ and $\mu \in \mathcal{M}$ an optimal transport for (OT). By Nutz (2022, Lemma 5.3), there exists $\mu^\eta \in \mathcal{M}$ such that

$$\left| \int c \mathrm{d}\mu^\eta - \int c \mathrm{d}\mu \right| \le \eta.$$

Note that $H(\mu^\eta) < \infty$. Then applying Lemma D.5 yields the conclusion. ☐

Then by Lemma D.4, the conclusion of Proposition 3.2 follows.

Appendix E. Optimality in semiparametric models

We generalize the setup presented in the main texts to allow more flexible sampling distributions for observable data. Our setup here basically follows Christensen *et al.* (2025, Section 5). Assume that data $Z^n = (Z_1, \ldots, Z_n)$ are i.i.d. and Z_i follows the distribution $P_{\theta,\eta}$ indexed by $\theta \in \Theta \subset \mathbb{R}^k$ and $\eta \in \mathcal{H}$, where η is a possibly infinite-dimensional nuisance parameter. For instance, in a GMM model, $\mathcal{H}$ is the set of marginal distributions η of Z_i where for each $\eta \in \mathcal{H}$, there exists $\theta \in \Theta$ such that η satisfies the moment restriction $\int g(\theta, z)\mathrm{d}\eta = 0$, given some known vector function g.

It is said that $\mathcal{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in \mathcal{H}\}$ has the *least favorable submodels* at (θ, η) if there exist an open neighborhood $\Theta_{\theta,\eta}$ of θ and a map $\Theta_{\theta,\eta} \ni t \mapsto \eta_t \in \mathcal{H}$ such that the parametric submodel $\{P_{t,\eta_t} : t \in \Theta_{\theta,\eta}\}$ has the density function p_{t,η_t} with respect to a common dominating measure ν and satisfies the DQM condition

$$\int \left[\sqrt{p_{\theta+h,\eta_\theta}} - \sqrt{p_{\theta,\eta_\theta}} - \frac{1}{2} h^\top \dot{\ell}_{\theta,\eta} p_{\theta,\eta_\theta} \right]^2 \mathrm{d}\nu = o(\|h\|^2), \quad \text{as } h \to 0,$$

where $\dot{\ell}_{\theta,\eta} : \mathcal{Z}^n \to \mathbb{R}^k$ is the efficient score function for θ. Thus, the parametric submodel $\{P_{t,\eta_t} : t \in \Theta_{\theta,\eta}\}$ achieves the semiparametric efficiency bound by the inverse of $I_{\theta,\eta} := \int \dot{\ell}_{\theta,\eta} \dot{\ell}_{\theta,\eta}^\top \mathrm{d}P_{\theta,\eta_\theta}$. For each (θ, η), the least favorable submodels need not to be unique. Picking one of them gives no loss of generality because they all behave in the same manner asymptotically.

Following the parametric model, we assume that the planner's utility function w only depends on θ, and not on the nuisance parameter η.

E.1. Decision theoretic framework and rules.

Fix $(\theta_0, \eta_0) \in \Theta \times \mathcal{H}$. Consider a least favorable submodel $\{P_{\beta(t)} : t \in \Theta_{\theta_0,\eta_0}\}$, where $\beta(t) = (t, \eta_t)$. Under the reparametrization $t = \theta_0 + h/\sqrt{n} = \theta_{nh}$, Z^n follows the distribution $P^n_{\beta(\theta_{nh})}$. We denote $\overset{h}{\rightsquigarrow}$ by the weak convergence along the path $P^n_{\beta(\theta_{nh})}$, $\overset{h}{\to}$ by the convergence in probability along $P^n_{\beta(\theta_{nh})}$, and $\overset{0}{\to}$ by the convergence in probability along $P^n_{\theta_0,\eta_0}$.

We define the class of the sequences of rules by

$$\mathcal{D} := \left\{ \{\mu_n\} : \mu_n(Z^n) \overset{h}{\rightsquigarrow} Q_{\theta_0,h} \text{ and } \sqrt{n} P^n_{\beta(\theta_{nh})}(\mu_n \in A_0) \to 0 \text{ as } n \to \infty \quad \forall h \in \mathbb{R}^k, \forall \theta_0 \in \Theta \right\}.$$

The optimality criterion in this semiparametric model is based on the least favorable submodels. The *risk* associated with the map $Z^n \mapsto \mu(Z^n) \in \mathcal{M}$ at $(\theta, \eta) \in \Theta \times \mathcal{H}$ is given by

$$R(\mu, (\theta, \eta)) := \mathbb{E}_{P^n_{\theta,\eta}} \left[W^*_{\mathcal{M}}(\theta) - W(\theta, \mu(Z^n)) \right],$$

where the expectation is taken with respect to the sampling distribution $P^n_{\theta,\eta}$ of Z^n. Let π be any prior density function on Θ that is continuous and positive at θ_0. Then, a sequence of rules $\{\mu^*_n\} \in \mathcal{D}$ is said to be *(semiparametrically) average optimal* if $\{\mu^*_n\}$ attains the infimum of the asymptotic risk function:

$$\inf_{\{\mu_n\} \in \mathcal{D}} \liminf_{n \to \infty} \int \sqrt{n} R(\mu_n, \beta(\theta_{nh})) \pi(\theta_{nh}) \mathrm{d}h. \tag{E.1}$$

E.2. **Quasi-Bayesian implementation of the Bayesian rules.** We replace a posterior function specified in the parametric setup by a quasi-posterior. Let $\hat{\theta}_n$ be the (semiparametrically) efficient estimator of θ, and $\hat{I}_n^{-1}$ be a consistent estimator of the asymptotic covariance $I^{-1}_{\theta,\eta}$. We combine the limited-information quasi-likelihood $N(\hat{\theta}_n, (n\hat{I}_n)^{-1})$ for θ and a prior π on Θ to obtain the quasi-posterior

$$\pi_n(\theta | Z^n) \propto \exp\left(-\frac{1}{2} (\theta - \hat{\theta}_n)^\top (n\hat{I}_n)(\theta - \hat{\theta}_n) \right) \pi(\theta).$$

We compute the Bayesian rules using $\pi_n(\theta|Z^n)$; i.e.,

$$\mu^B_n(z) \in \mathcal{M}_{opt}(z) := \arg\max_{\mu \in \mathcal{M}} \int \sqrt{n} W(\theta, \mu) \mathrm{d}\pi_n(\theta|z).$$

Following the parametric model, we construct a unique $\{\mu^B_n(z)\}$ where $\mu^B_n(z)$ minimizes the penalty function H over $\mathcal{M}_{opt}(z)$.

E.3. **Optimality results.** As an analog for Assumption 3.1 in the parametric models, we impose the following assumptions.

Assumption E.1. *(i)* Θ *is open.*

(ii) $\mathcal{P}$ *has a least favorable submodel at each* $(\theta_0, \eta_0) \in \Theta \times \mathcal{H}$.

(iii) I_{θ_0,η_0} *is finite and nonsingular at each* $(\theta_0, \eta_0) \in \Theta \times \mathcal{H}$.

(iv) For each $(\theta_0, \eta_0) \in \Theta \times \mathcal{H}$ *and each* $h \in \mathbb{R}^k$, *(iv-a)* $\sqrt{n} P^n_{\beta(\theta_{nh})}(\|\hat{\theta}_n - \theta_0\| > \varepsilon) \to 0$ *for each* $\varepsilon > 0$ *as* $n \to \infty$, *and (iv-b) there exists* $c \in (0,1)$ *such that* $\sqrt{n} P^n_{\beta(\theta_{nh})}(c \leq \hat{\lambda}_{\min}, \hat{\lambda}_{\max} \leq c^{-1}) \to 0$ *as* $n \to \infty$,

(v) For each $(\theta_0, \eta_0) \in \Theta \times \mathcal{H}$, *(v-a)* $\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{h}{\rightsquigarrow} Z$ *with* $Z \sim N(h, I^{-1}_{\theta_0,\eta_0})$ *for all* $h \in \mathbb{R}^k$ *as* $n \to \infty$, *and (v-b)* $\hat{I}_n \overset{0}{\to} I_{\theta_0,\eta_0}$ *as* $n \to \infty$.

Theorem E.1. *Under Assumptions E.1 and 3.2–3.6,* $\{\mu^B_n\} \in \mathcal{D}$ *is average optimal.*

Proof. Once we fix the parameter $(\theta_0, \eta_0) \in \Theta \times \mathcal{H}$, the least favorable submodel $\{P_{\beta(t)} : t \in \Theta_{\theta_0,\eta_0}\}$ becomes a parametric model. Hence, only slight modifications from the proof of Theorem 3.3 are needed. Specifically, Lemmas A.2 and A.6 follow in the same manner. For Lemma A.3, we need a modification to show $\sqrt{n} P^n_{\beta(\theta_{nh})}(\mu^B_n(Z^n) \notin A_0) \to 0$ for all $h \in \mathbb{R}^k$, which is given in Lemma E.2

below. For Lemmas A.7 and A.8, we need to use the quasi-posterior counterparts of the Bernstein-von Mises theorem given by Christensen *et al.* (2025, Lemma 5) and Proposition A.1 given by Xu (2024, Lemma A.5). Auxiliary lemmas given in Appendix B do not need modifications. □

Lemma E.2. *The Bayesian rule* $\{\mu_n^B(Z^n)\}$ *satisfies* $\sqrt{n}P^n_{\beta(\theta_{nh})}(\mu_n^B(Z^n) \notin A_0) \to 0$ *as* $n \to \infty$.

Proof. From Lemma A.3 (i), for any $\theta_0 \in \Theta$, there are $\bar{n}$ and $\varepsilon'_n(\bar{n})$ (which is at the order of $n^{\alpha+1}$ for some $\alpha \geq 1$) such that for all $n \geq \bar{n}$,

$$P^n_{\theta_{nh}}\left(\mu_n^B(z) \notin A_0\right) \leq P^n_{\theta_{nh}}\left(\pi_n\left(N_{1/n}(\theta_0)^c|z\right) > 2\varepsilon'_n\right)$$

Under Assumption E.1 (iii) and (iv), Christensen *et al.* (2025, Lemma 12) implies that

$$\sqrt{n}P_{\beta(\theta_{nh})}\left(\pi_n\left(N_{1/n}(\theta_0)^c\right) > 2\varepsilon'_n\right) \to 0$$

as $n \to \infty$. □

References

Adjaho, C. and T. Christensen (2023). "Externally valid policy choice." arXiv: 2205.05561 [econ.EM].

Angrist, J., E. Bettinger, and M. Kremer (2006). "Long-term educational consequences of secondary school vouchers: Evidence from administrative records in Colombia." *American Economic Review* 96.3, 847–862. DOI: 10.1257/aer.96.3.847.

Angrist, J., E. Bettinger, and M. Kremer (2019). "Replication data for: Long-term educational consequences of secondary school vouchers: Evidence from administrative records in colombia." *American Economic Association*. DOI: 10.3886/E116225V1.

Aradillas Fernández, A., J. L. Montiel Olea, C. Qiu, J. Stoye, and S. Tinda (2024). "Robust Bayes treatment choice with partial identification." arXiv: 2408.11621 [econ.EM].

Athey, S. and S. Wager (2021). "Policy learning with observational data." *Econometrica* 89.1, 133–161. DOI: 10.3982/ECTA15732.

Banerjee, A. V., S. Chassang, S. Montero, and E. Snowberg (2020). "A theory of experimenters: Robustness, randomization, and balance." *American Economic Review* 110.4, 1206–1230. DOI: 10.1257/aer.20171634.

Bhattacharya, D. and P. Dupas (2012). "Inferring welfare maximizing treatment assignment under budget constraints." *Journal of Econometrics* 167.1, 168–196. DOI: 10.1016/j.jeconom.2011.11.007.

Chen, H. and P. Guggenberger (2025). "A note on minimax regret rules with multiple treatments in finite samples." *Econometric Theory*, 1–27. DOI: 10.1017/S0266466625000040.

Chewi, S., J. Niles-Weed, and P. Rigollet (2024). "Statistical optimal transport." arXiv: 2407.18163 [math.ST].

Christensen, T., H. R. Moon, and F. Schorfheide (2025). "Optimal decision rules when payoffs are partially identified." arXiv: 2204.11748v2 [econ.EM].

Clarke, B. S. and A. R. Barron (1990). "Information-theoretic asymptotics of Bayes methods." *IEEE Transactions on Information Theory* 36.3, 453–471. DOI: 10.1109/18.54897.

Dehejia, R. H. (2005). "Program evaluation as a decision problem." *Journal of Econometrics* 125.1-2, 141–173. DOI: 10.1016/j.jeconom.2004.04.006.

Fang, Z. and A. Santos (2019). "Inference on directionally differentiable functions." *Review of Economic Studies* 86.1, 377–412. DOI: 10.1093/restud/rdy049.

Flamary, R., N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer (2021). "POT: Python optimal transport." *Journal of Machine Learning Research* 22.78, 1–8.

Ghosal, S. and A. W. van der Vaart (2017). *Fundamentals of Nonparametric Bayesian Inference*. Vol. 44. Cambridge University Press.

Ghosh, J. K. and R. V. Ramamoorthi (2003). *Bayesian Nonparametrics*. Springer New York.

Giacomini, R. and T. Kitagawa (2021). "Robust Bayesian inference for set-identified models." *Econometrica* 89.4, 1519–1556. DOI: 10.3982/ECTA16773.

Gilboa, I. and D. Schmeidler (1989). "Maxmin expected utility with non-unique prior." *Journal of Mathematical Economics* 18.2, 141–153. DOI: 10.1016/0304-4068(89)90018-9.

Guggenberger, P., N. Mehta, and N. Pavlov (2024). "Minimax regret treatment rules with finite samples when a quantile is the object of interest." *Working Paper*.

Hazard, Y. and T. Kitagawa (2025). "Who with whom? Learning optimal matching policies." arXiv: 2507.13567 [econ].

Heinonen, J. (2001). *Lectures on Analysis on Metric Spaces*. Springer New York, NY.

Hirano, K. and J. R. Porter (2009). "Asymptotics for statistical treatment rules." *Econometrica* 77.5, 1683–1701. DOI: 10.3982/ECTA6630.

Ishihara, T. and T. Kitagawa (2024). "Evidence aggregation for treatment choice." arXiv: 2108.06473v2 [econ.EM].

Kido, D. (2022). "Distributionally robust policy learning with wasserstein distance." arXiv: 2205.04637 [econ].

Kido, D. (2023). "Locally asymptotically minimax statistical treatment rules under partial identification." arXiv: 2311.08958 [econ.EM].

Kim, J.-Y. (2002). "Limited information likelihood and Bayesian analysis." *Journal of Econometrics* 107.1-2, 175–193. DOI: 10.1016/S0304-4076(01)00119-1.

Kitagawa, T., S. Lee, and C. Qiu (2024). "Treatment choice with nonlinear regret." arXiv: 2205.08586 [econ.EM].

Kitagawa, T. and A. Tetenov (2018). "Who should be treated? Empirical welfare maximization methods for treatment choice." *Econometrica* 86.2, 591–616. DOI: 10.3982/ECTA13288.

Manski, C. F. (2004). "Statistical treatment rules for heterogeneous populations." *Econometrica* 72.4, 1221–1246. DOI: 10.1111/j.1468-0262.2004.00530.x.

Manski, C. F. (2007). "Minimax-regret treatment choice with missing outcome data." *Journal of Econometrics* 139.1, 105–115. DOI: 10.1016/j.jeconom.2006.06.006.

Masten, M. A. (2023). "Minimax-regret treatment rules with many treatments." *Japanese Economic Review* 74.4, 501–537. DOI: 10.1007/s42973-023-00147-0.

Mbakop, E. and M. Tabord-Meehan (2021). "Model selection for treatment choice: Penalized welfare maximization." *Econometrica* 89.2, 825–848. DOI: 10.3982/ECTA16437.

Montiel Olea, J. L., C. Qiu, and J. Stoye (2024). "Decision theory for treatment choice problems with partial identification." arXiv: 2312.17623 [econ].

Nutz, M. (2022). "Introduction to entropic optimal transport."

Qi, Z., J.-S. Pang, and Y. Liu (2023). "On robustness of individualized decision rules." *Journal of the American Statistical Association* 118.543, 2143–2157. DOI: 10.1080/01621459.2022.2038180.

Römisch, W. (2004). "Delta method, infinite dimensional." *Encyclopedia of Statistical Sciences* 3.

Shapiro, A. (1990). "On concepts of directional differentiability." *Journal of Optimization Theory and Applications* 66, 477–487.

Si, N., F. Zhang, Z. Zhou, and J. Blanchet (2020). "Distributionally robust policy evaluation and learning in offline contextual bandits." In: *International Conference on Machine Learning*. PMLR, 8884–8894.

Stoye, J. (2009). "Minimax regret treatment choice with finite samples." *Journal of Econometrics* 151.1, 70–81. DOI: 10.1016/j.jeconom.2009.02.013.

Stoye, J. (2012). "Minimax regret treatment choice with covariates or with limited validity of experiments." *Journal of Econometrics* 166.1, 138–156. DOI: 10.1016/j.jeconom.2011.06.012.

Sun, L. (2024). "Empirical welfare maximization with constraints." arXiv: 2103.15298 [econ.EM].

Tetenov, A. (2012). "Statistical treatment choice based on asymmetric minimax regret criteria." *Journal of Econometrics* 166.1, 157–165. DOI: 10.1016/j.jeconom.2011.06.013.

van der Vaart, A. W. (1991). "An asymptotic representation theorem." *International Statistical Review* 59.1, 97–121. DOI: 10.2307/1403577.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Vol. 3. Cambridge University Press.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer-Verlag. DOI: 10.1007/978-1-4757-2545-2.

Villani, C. (2009). *Optimal Transport: Old and New*. Vol. 338. Springer.

Xu, H. (2024). "Asymptotic analysis of point decisions with general loss functions." *Working Paper*.

Yata, K. (2023). "Optimal decision rules under partial identification." arXiv: 2111.04926 [econ.EM].

Department of Economics, University of Rochester, Rochester, NY 14627, USA.
Email address: ksunada@ur.rochester.edu

Department of Economics, University of Rochester, Rochester, NY 14627, USA.
Email address: kizumi3@ur.rochester.edu