Policy Learning for Localized Interventions from Observational Data

Myrl G. Marmarelis **Fred Morstatter** **Aram Galstyan** **Greg Ver Steeg**
USC Information Sciences Institute
4676 Admiralty Way, Marina del Rey, CA 90292

Abstract

A largely unaddressed problem in causal inference is that of learning reliable policies in continuous, high-dimensional treatment variables from observational data. Especially in the presence of strong confounding, it can be infeasible to learn the entire heterogeneous response surface from treatment to outcome. It is also not particularly useful, when there are practical constraints on the size of the interventions altering the observational treatments. Since it tends to be easier to learn the outcome for treatments near existing observations, we propose a new framework for evaluating and optimizing the effect of small, tailored, and localized interventions that nudge the observed treatment assignments. Our doubly robust effect estimator plugs into a policy learner that stays within the interventional scope by optimal transport. Consequently, the error of the total policy effect is restricted to prediction errors nearby the observational distribution, rather than the whole response surface.

1 INTRODUCTION

Improvements in predictive power from large statistical models do not always translate to better decision making. The best way to support decision-making is to infer the outcomes from possibly relevant interventions. Models built to describe observations like in supervised or self-supervised learning tasks are not always adequate to predict interventional outcomes. In the field of causal inference, the theory of causal-effect estimation in a *potential outcomes* framework is largely concerned with building statistical models that can predict interventional outcomes explicitly.

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

From observational data, it can be useful to learn the causal effects of a treatment variable and then construct a policy for prescribing treatments with the goal of maximizing overall outcomes (Athey and Wager, 2021). This problem is often called offline policy learning because one cannot take actions and observe new outcomes like in an online setting. The structure of the causal system, and in particular that of the treatment variable, can have severe implications on the feasibility of learning optimal policies. Treatments are commonly assumed to be binary, although more flexible settings are gaining traction, including for discrete multi-valued treatments (Zhou et al., 2023; Kallus et al., 2022; Uehara et al., 2020), continuous treatments (Demirer et al., 2019), and discrete multivariate treatments (Liang et al., 2018; Xu et al., 2023b,a).

When treatments are continuous and possibly multivariate, it can be quite difficult to learn the full response surface of every unit to every treatment value. Furthermore, prevailing estimators require learning the conditional probability density of the observational treatment propensity, e.g. Nie et al. (2021); Colangelo and Lee (2020); Marmarelis et al. (2023); Kallus and Zhou (2018). That task might not scale well with covariate shift and increasing treatment dimensionality.

It could be simpler to estimate local causal derivatives (Hines et al., 2023; Chernozhukov et al., 2022). Focusing on derivatives restricts the learning problem to the parts of the response surface near existing observations. Causal derivatives at observed treatments are informative of incremental effects from small interventions. Small interventions are often the most achievable. Despite these benefits, it is not always obvious how to learn a policy from causal derivatives. How small of an intervention is small enough? How far can a derivative extrapolate? Never mind the additional considerations for outcome predictors with well-behaved derivatives, which are needed in estimators of causal derivatives.

We therefore propose to learn the effects of *nudging*[1] the treatment variable (§2), and to optimize budget-constrained policies thereof (§3). A causal-effect estimator specifically for nudges coming from a *nudge*

prior, representing the interventional scope and perhaps budget under consideration, facilitates reliable learning of nudge policies (§5). Figure 1 shows a simple illustrative scenario calling for nudges.

Our solution involves a few novelties. We formulate a learning objective (§2.3) for directly debiasing nudge-effect estimates in a doubly robust framework. While developing the policy learner, we discover a connection to optimal transport with an unorthodox cost function (§3.2), on transferring nudges to observational units. This link further reveals the possibility of an efficient, information-bottlenecked solver (§3.3).

Example (wildfires). Suppose we are attempting to spatially target costly interventions that could reduce the proclivity for wildfires. We are studying satellite images (Drusch et al., 2012) to identify the pixels most conducive to intervention before summer. From records of past summers and perhaps large-scale climate models (Rodgers et al., 2021), we have learned a spatial forecasting model for wildfire occurrence given a satellite image of surrounding vegetation, moisture, and topography. If the intervention to target is a reduction in vegetation, then the machine-learning problem could be to identify the high-vegetation pixels that would, once lowered, *causally* reduce the risk of a wildfire. We are effectively *nudging* multiple continuous-valued spatial pixels because we are aiming to reduce vegetation to varying degrees as determined by the optimal policy.

Relation to reinforcement learning. In the literature of proximal policy optimization (PPO) (Schulman et al., 2015, 2017; Ouyang et al., 2022), a policy model is optimized with respect to some reward model within a trust region. Usually, the trust region is defined by a cutoff or penalty on the KL-divergence between the old (original or logging) and new (learned) policy. This prevents domain shift due to the updated policy moving far from the data-generating process. Our conception of nudge priors can be interpreted as a trust region as well. However, to be faithful to PPO, we would have the learned policy be similar to the treatment propensity—the conditional distribution governing treatment assignments. The true propensity is unknown and has to be estimated in our problem setup, and it can be difficult to guarantee its accuracy. It can be untenable to base the trust region on this estimate. Instead, we use a nudge *prior* that fixes the *marginal* distribution of nudge policies over the data. Our lack of a known logging policy also separates us from offline contextual bandits (Yang et al., 2023).

[1]Even though we believe a "nudge" is the best term for a small intervention, somewhere between infinitesimal and global, we wish *not* to encourage associations with nudge theory from the social/behavioural sciences (Ewert, 2020).

Doubly robust estimation. Stemming from semiparametric estimation theory (Hines et al., 2022; Kennedy, 2022), the framework of *double machine learning* (Chernozhukov et al., 2018) has become immensely useful in the state of the art in causal-effect estimation and offline policy learning (e.g. Oprescu et al., 2023; Zhou et al., 2023; Kallus et al., 2022). We employ this theory to derive estimators with increased robustness to misspecification in the learned models.

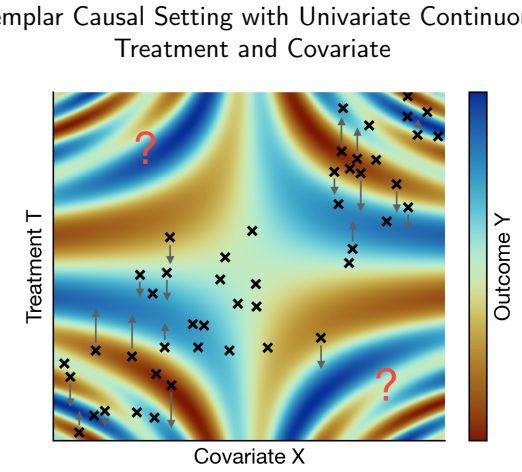

Figure 1. With continuous-valued treatments and significant covariate shift illustrated by the observations marked by **x**s, it can be infeasible to learn the full response surface $T, X \mapsto Y$. In this example the upper left and lower right corners of the response would be impossible to predict without significant prior knowledge. Subsequently, it would be impractical to learn policies $X \mapsto T$ over the full surface, and perhaps infeasible to act on them. We propose to learn localized policies shown by the nudging arrows in gray.

2 ESTIMATING NUDGE EFFECTS

The first step of our methodology is to estimate nudge effects. The overall approach is to use generic machine learning to estimate certain functions of the data, which are combined to form robust causal estimates.

2.1 Problem Setup

We consider vector treatments with unbounded support $T \in \mathcal{T} = \mathbb{R}^{d_T}$ under otherwise standard potential-outcomes assumptions listed in Assumption 1 on the system of *(outcome, treatment, covariate)* variables $W = (Y, T, X)$. Our aim is to learn the effect of an interventional treatment nudge $\xi \in \mathbb{R}^{d_T}$ concretely defined as $\Delta_\xi Y \triangleq Y(T+\xi) - Y(T)$, where $Y(t)$ is the potential outcome at t. Only the realized outcome Y at the assigned treatment T is ever observed for each

unit in the sample. The rest must be inferred in a way that accounts for selection and confounding biases.

Assumption 1 (Potential Outcomes)**.** The standard causal setting due to Rubin (1974).

(a) Stable unit treatment value assumption (SUTVA). The potential outcomes for a unit do not depend on the treatment assigned to any other unit.

(b) Overlap/positivity. All treatment values have a nonzero probability of occurring for every unit.

(c) Ignorability. Potential outcomes are independent of the treatment after conditioning on covariates.

Assumptions for nudges. A core assumption for the learning problem posed in this work is tied to the nudges under consideration. Our prior knowledge in this regard will be encapsulated by a distribution over the nudges, with a probability density $p(\xi)$. This *nudge prior* governs the domain for the estimation problem covered in this section, for the policy learning in §3, as well as the error bounds in §5. The distributional assumption subsumes possible budgetary requirements that could be written as expectations over nudges, like a threshold over the average magnitude of prescriptions to a sample. It can also reflect our degree of confidence in the learned models, depending on the problem difficulty. Estimates that cannot extrapolate far should only be trusted with narrow nudge priors.

2.2 Our Approach

We outline an estimator that synergizes two predictive models for nuisance parameters, so-called because they parametrize the causal-outcome estimates. These will be combined to form a robust estimate of the nudge effect $\Delta_\xi Y$ conditional to other observable features from X. We learn the following models:

- conditional outcome $\mu(T, X) \triangleq \mathbb{E}[Y \mid T, X]$, and
- propensity ratio $\eta_\xi(T, X) \triangleq f_\xi(T|X) \,/\, f_0(T|X)$.

The propensity density function is denoted as $f_0(T|X)$, belonging to a family of propensities that have been shifted by a nudge ξ, defined as $f_\xi(t|X) = f_0(t - \xi|X)$ induced by the transport map $(T, X) \mapsto (T + \xi, X)$.

Taking inspiration from a rich line of work on efficient influence functions (Hines et al., 2022; Kennedy, 2022), we propose a *pseudo-outcome* for nudge effects that combines nuisances and data:

$$\varphi_\xi \triangleq [\eta_\xi(T, X) - 1]\,[Y - \mu(T, X)] + \Delta_\xi \mu(T, X), \quad (1)$$

where $\Delta_\xi \mu(T, X) \triangleq \mu(T + \xi, X) - \mu(T, X)$. Like other parameters studied in the double machine learning literature, this pseudo-outcome is an unbiased estimator for the nudge effect even if only one of the two nuisance estimates $\{\hat{\mu}(T, X), \hat{\eta}_\xi(T, X)\}$ is correctly specified (see §5). Hence, $\hat{\varphi}_\xi$ is considered doubly robust/debiased. In fact, by iterated expectation over Y and T, it can be shown that $\mathbb{E}[\hat{\varphi}_\xi \mid X]$ is a doubly robust estimator of the heterogeneous nudge effects, $\mathbb{E}[\Delta_\xi Y \mid X]$. As is explored further in §5, the local error of $\hat{\varphi}_\xi$ at a specific X is the product of the errors of the two nuisance estimates $\{\hat{\mu}(T, X), \hat{\eta}_\xi(T, X)\}$, enabling faster convergence for the effect estimate. We state a proposition that serves as a theoretical basis for deriving Equation 1;

Proposition 1. *The efficient influence function for the average nudge effect is* $\varphi_\xi - \mathbb{E}[\Delta_\xi Y]$.

Kennedy (2020), Oprescu et al. (2023), and others have already proposed regression on pseudo-outcomes with a data-splitting strategy in order to accurately estimate conditional average treatment effects (CATEs) for *binary* treatments. In a typical data-splitting procedure, one estimates the functions (nuisance parameters) $\{\mu(T, X), \eta_\xi(T, X)\}$ in one data partition, and then computes pseudo-outcome estimates $\hat{\varphi}_\xi(Y, T, X)$ on another data partition using $\{\hat{\mu}(T, X), \hat{\eta}_\xi(T, X)\}$. Heterogeneous effects can be identified by regressing on the pseudo-outcomes in the second data partition.

2.3 Estimating the Propensity Ratio

The solution we present includes an approximation for the propensity-density ratio parameter η_ξ. Its form is motivated by its limiting behavior as the negative logarithmic gradient of the propensity density. Keeping in mind that the estimate $\hat{\eta}_\xi$ should be reliable in the regime of the nudge prior $p(\xi)$, we chose to construct a set of learning problems defined over a sample of nudges. We begin with a first-order multivariate Taylor expansion of $\log f_\xi(T|X)$ in ξ,

$$\log f_\xi(T|X) = \log f_0(T|X) + \xi \cdot g(T, X) + \mathcal{O}(\|\xi\|^2)$$

where the identity $f_\xi(T|X) = f_0(T - \xi|X)$ reveals that $g(T, X)$ is indeed $-\nabla_T \log f_0(T|X)$. This expansion suggests an approximation to the propensity ratio,

$$\eta_\xi(T, X) = \frac{f_\xi(T|X)}{f_0(T|X)} \approx \exp\{\xi \cdot g(T, X)\},$$

which leads to a reparametrization of the estimate $\hat{\eta}_\xi(T, X)$ in $\hat{g}(T, X)$ for any ξ drawn from a prior of relatively small nudges. Even though $g(T, X)$ is defined as a logarithmic gradient, the optimal estimator for $\hat{g}(T, X)$ over $p(\xi)$ would minimize an average loss over the nudges. This distinguishes our solution from traditional score matching (Hyvärinen and Dayan, 2005).

Probabilistic classification (PC). Let $q_\xi(T, X)$ be the probability of T coming from $f_\xi(T|X)$ rather than $f_0(T|X)$. Then it follows that

$$\eta_\xi(T, X) = \frac{q_\xi(T|X)}{1 - q_\xi(T|X)},$$
$$\log \eta_\xi(T, X) = \log \frac{q_\xi(T|X)}{1 - q_\xi(T|X)} = \xi \cdot g(T, X).$$

Hence $\xi \cdot g(T, X)$ gives classification logits for $f_\xi(T|X)$ versus $f_0(T|X)$. In this way, we propose to learn $\hat{g}(T, X)$ from a set of classification problems over a nudge sample by learning to classify $(T + \xi, X)$ as positive and (T, X) as negative over the data (Y, T, X) and the nudge prior $p(\xi)$. Probablistic classification is a simple approach to estimating density ratios (Tibshirani et al., 2019) and our scheme of formulating many small classification problems is partly inspired by recent work like that of Choi et al. (2022). All in all, our loss function for $\hat{g}(t, x)$ is given as

$$\mathcal{L}^{\mathrm{PC}}[g] \triangleq \hat{\mathbb{E}}_{(T,X)\times p(\xi)} \Big[\log \sigma(\xi \cdot g(T, X)) - \log \sigma(\xi \cdot g(T + \xi, X)) \Big], \quad (2)$$

where $\sigma(\cdot)$ is the logistic sigmoid. As already mentioned, this nuisance parameter $\hat{g}(T, X)$ is estimated on one dataset split and then used on the other dataset split. When $\hat{g}(T, X)$ is parametrized as a deep neural network, we found that it can be helpful to calibrate it on the second split with a low-dimensional adjustment like temperature scaling (Guo et al., 2017). Well-calibrated and smoothed logits are paramount to the stability of probabilistic classification.

Comparison to denoising score matching (SMD). The recent success of diffusion models (Croitoru et al., 2023) has underscored the popularity of SMD (Vincent, 2011; Swersky et al., 2011). The objective of SMD can be understood as score matching on a smoothed (noised) version of the data distribution. It is similar to our PC approach in that the score, i.e. log-gradient of a density, is related to the log-ratio of an infinitesimally perturbed density against the original density. See the remark on causal derivatives, Equation 4. The smoothing brings enjoyable finite-sample qualities and also simplifies the learning objective. If we were to apply SMD to estimating $g(T, X)$ using the existing domain knowledge for nudges, then the smoothing probabilistic kernel over treatments would be defined as additive noise in terms of nudges, probably coming from $p(\xi)$. Therefore the SMD loss would be

$$\mathcal{L}^{\mathrm{SMD}}[g] \triangleq \hat{\mathbb{E}}_{(T,X)\times p(\xi)} \, \|\nabla \log p(\xi) - g(T + \xi, X)\|_2^2 \quad (3)$$

and it would uncover $\hat{g}(t, x) \approx \nabla \log \mathbb{E}_{p(\xi)} f_0(t - \xi \mid x)$. On the other hand, our PC objective (Equation 2) would uncover local linear coefficients $\hat{g}$ such that $\xi \cdot \hat{g}(t, x) \approx \log f_0(t - \xi \mid x)/f_0(t \mid x)$, trained by a logistic cross-entropy loss. While SMD could certainly be used to approximate our learning task, it would be suboptimal especially because the nudge expectation resides within the logarithm.

Remark (causal derivative). Letting $\xi = \varepsilon v$ and ∇_v denote a directional derivative in the first argument of a function with unit vector v, the quantity

$$\varepsilon^{-1}[\eta_\xi(T, X) - 1] \xrightarrow[\varepsilon \to 0]{} -\nabla_v \log f_0(T|X) \quad (4)$$

recovers a multivariate version of the Riesz representer for the average causal derivative (ACD) as considered in prior works like Chernozhukov et al. (2022). The form of Equation 1 resembles a finite-difference version of the pseudo-outcomes for the ACD.

3 LEARNING NUDGE POLICIES

Our goal is to make policy prescriptions for nudge interventions that reliably maximize nudge effects in expectation. The policy should adhere to the nudge prior for two reasons: first, any interventional budget constraints should be respected. Second, as the effect estimates are calibrated for the nudge prior distribution, the policy should not stray from that domain where generalization degrades. With that in mind, we seek a learned heterogeneous policy that is a function of some flexibly defined variable $U \in \mathcal{U}$ that satisfies $U = f(X)$ for some f, allowing for the simplest case $U = X$, or for the cases of U being a subset of the covariate features when one wishes to generalize to a broader population with fewer recorded attributes. Consider a scenario where one learns from a detailed survey (using X) and then makes broader prescriptions (on a simpler U).

Letting Π be the set of conditional density functions in $P(\Xi|\mathcal{U})$, the ideal nudge policy is characterized by an optimization problem constrained on the marginal.

Definition 1 (Optimal Nudge Policy)**.** A nudge policy given as a probability density $\pi(\xi|U)$ conditional on features U is considered optimal if it solves the following constrained optimization problem:

$$\max_{\pi \in \Pi} \; \mathbb{E}\Big[\int_\Xi \Delta_\xi Y \, \pi(\xi|U) \, \mathrm{d}\xi \Big],$$
$$\text{s.t.} \quad \forall \xi \in \Xi, \; \mathbb{E}[\pi(\xi|U)] = p(\xi).$$

The attained maximum value is considered the *nudge-policy effect* because it integrates the population's nudge effects over the policy prescriptions.

Note on notation. As in the other sections of this paper, the bare expectations $\mathbb{E}$ are with respect to the data $W = (Y, T, X)$, and the training set for the

nuisances, but not the nudge variable ξ. Expectations over the nudge *only* are denoted as $\mathbb{E}_\pi$ or the integral form shown in Definition 1. Of course, the nudge domain Ξ is equal to the treatment domain $\mathcal{T}$, and they will be used interchangeably.

The infinite and idealized optimization problem in Definition 1 is intractable per se. To proceed, we introduce two approximate solution schemes denoted as `IB` and `OT`, eventually synergizing into `IB+OT` from the desirable aspects of both perspectives.

3.1 `IB` — *Information Bottleneck* Policy Learner

The first approach is to solve a more constrained optimization problem inspired by the information-bottleneck principle (Tishby et al., 2000) that is commonly employed in variational bounds for intractable problems in representation learning (Higgins et al., 2016). Notice that the hard constraint on the marginal can be equivalently stated in the Kullback-Leibler (KL) divergence, or relative entropy:

$$\mathrm{D}\big(\ \mathbb{E}[\pi(\xi|U)]\ \|\ p(\xi)\ \big) = 0. \tag{5}$$

This can be turned into a single soft constraint by penalizing the objective function with a Lagrange multiplier. However, it is still difficult to compute over a large sample. We turn to the mutual information between ξ and X induced by a given policy π,

$$\begin{aligned} I_\pi(\xi;U) &= \mathrm{D}\big(\ \pi(\xi|U)p(U)\ \|\ p(\xi)p(U)\ \big) \\ &= \mathbb{E}[\ \mathrm{D}\big(\ \pi(\xi|U)\ \|\ p(\xi)\ \big)\ ], \end{aligned} \tag{6}$$

which implicitly asserts that the nudge policy's marginal is identical to the nudge prior. It is easier to compute because if the policy and prior are of the same parametric family (e.g. conditionally Gaussian), the inner divergence quantity can be solved analytically. By Jensen's inequality, this mutual information is an upper bound on the marginal divergence constraint of Equation 5. Therefore, the mutual information acts as a stronger constraint: not only does it enforce correct marginal behavior, but it also limits the specificity of the policy. It introduces an information bottleneck. This additional regularization from Equation 6 on the policy implies that any policy prescription $\pi(\xi|u_i)$ for a particular unit u_i that strongly diverges from the nudge prior would only do so for a large apparent benefit. This phenomenon could be useful for ranking the most beneficial interventions. The `IB` policy-learning objective functional is stated as follows, for $\beta > 0$:

$$\mathcal{L}^{\texttt{IB}}[\pi,\beta] \triangleq \hat{\mathbb{E}}\big[\ -\mathbb{E}_\pi\,\hat{\varphi}_\xi + \beta\mathrm{D}\big(\ \pi(\xi|U)\ \|\ p(\xi)\ \big)\big]. \tag{7}$$

The objective $\mathcal{L}^{\texttt{IB}}$ can straightforwardly be used to train a neural network parametrizing $u \mapsto \pi(\xi|u)$ with hyperparameter β controlling the strength of the IB.

3.2 `OT` — *Optimal Transport* Policy Learner

The second approach is a natural consequence of the observation that the optimization problem in Definition 1 is an infinite linear program (Dantzig, 1963). Informally, we seek to maximize a massive sum of nudge effects weighted by the policy density function, subject to an equality constraint along every nudge value as well as constraints ensuring that the conditional policy is a valid probability density function.

This motivates a translation of Definition 1 to a form where all the constraints are explicit and the optimization is performed over all functions $\pi : \Xi \times \mathcal{X} \to \mathbb{R}$:

$$\begin{aligned} &\arg\max_\pi\ \mathbb{E}\Big[\int_\Xi \Delta_\xi Y\,\pi(\xi|U)\,\mathrm{d}\xi\Big], \\ &\text{s.t.}\quad \forall\xi,\ \mathbb{E}[\pi(\xi|U)] = p(\xi), && \textit{(nudge prior)} \\ &\qquad \forall u,\ \int_\Xi \pi(\xi|u)\,\mathrm{d}\xi = 1, && \textit{(policy validity \#1)} \\ &\qquad \forall(\xi,u),\ \pi(\xi|u) \geq 0. && \textit{(policy validity \#2)} \end{aligned}$$

The problem begins to look like an optimal transport with a transference plan $\pi'(\xi,u) = \pi(\xi|u)p(u)$ where $p(u)$ is the marginal density of X. The semblance becomes clearer when the second group of constraints is written as $\int_\Xi \pi'(\xi,u)\,\mathrm{d}\xi = p(u)$, mirroring the first group of constraints on $p(\xi)$. The connection between linear programming and optimal transport was first noticed by Kantorovich (1942). Concretely, the problem in Equation 8 is a Monge-Kantorovich optimal transport with cost function $-\Delta_\xi Y$, assuming it is continuous and finite (Villani et al., 2009). We substitute this ideal cost function with a pseudo-outcome approximation $c(\xi, W) \triangleq -\varphi_\xi(W)$; recall that $W = (Y,T,X)$ describes a whole observational unit.

$$\begin{aligned} \pi^\star = \arg\max_\pi\ &\mathbb{E}\Big[\int_\Xi \varphi_\xi\,\pi(\xi|U)\,\mathrm{d}\xi\Big], \\ \text{s.t.}\quad &\forall\xi,\ \mathbb{E}[\pi(\xi|U)] = p(\xi), \\ &\forall u,\ \int_\Xi \pi(\xi|u)\,\mathrm{d}\xi = 1, \quad \forall(\xi,u),\ \pi(\xi|u) \geq 0. \end{aligned} \tag{8}$$

We study the finite-sample version of Equation 8, which is again an optimal transport:

$$\begin{aligned} \hat{\pi}^{(n,m)} = \arg\max_\pi\ &\sum_{j=1}^{n}\sum_{i=1}^{m} \hat{\varphi}_{ij}\pi_{ij}, \\ \text{s.t.}\quad &\frac{1}{m}\sum_{i=1}^{m}\pi_{ij} = 1,\quad \frac{1}{n}\sum_{j=1}^{n}\pi_{ij} = 1,\quad \pi_{ij} \geq 0, \end{aligned} \tag{9}$$

where the quantities have been written in matrix form as $\hat{\varphi}_{ij} = \hat{\varphi}_{\xi_i}(w_j)$, $\pi_{ij} = \pi(\xi_i|u_j)/p(\xi_i)$ for convenience. We use i.i.d $\{\xi_i\}_{i=1}^m$ from the prior and $\{w_j = (y_j,t_j,x_j)\}_{j=1}^n$ from the second dataset split for computing $\hat{\varphi}_{ij}$ in $(m \times n)$ batches.

Even though we heretofore required that the policy marginal adhered to the nudge prior, in practice any additional sparsity in prescriptions for ineffective interventions is helpful. In other words, it is desirable to allow a policy to favor inaction. Any unmatched or not-fully-matched unit u_j to nudge points ξ_i implicitly matches it to the null nudge $\xi_0 = 0$ since $\Delta_{\xi=0} Y = 0$. We shall permit any amount of null mass to be allocated to units. Further, the problem naturally accommodates L^1-regularization of the policy by introducing a hyperparameter $\gamma \geq 0$. This would encourage policies to tend towards sparsity. Combining these two modifications to Equation 9 amounts to changing the equality constraints to inequalities and perturbing the objective:

$$\begin{aligned} \hat{\pi}_\gamma^{(n,m)} &= \arg\max_\pi \sum_{j=1}^{n}\sum_{i=1}^{m} (\hat{\varphi}_{ij} - \gamma)\pi_{ij}, \\ \text{s.t.} \quad & \sum_{i=1}^{m} \pi_{ij} \leq m, \quad \sum_{j=1}^{n} \pi_{ij} \leq n, \quad \pi_{ij} \geq 0. \end{aligned} \tag{10}$$

These finite linear programs can be solved straightforwardly on the second dataset split. A solution $[\hat{\pi}_{ij}]$ gives a particle approximation to the policy density for the observations by

$$\hat{\pi}(\xi | u_j) = \frac{1}{m} \sum_{i=1}^{m} \hat{\pi}_{ij}\, \delta(\xi - \xi_i).$$

Finally, we propose a synthesis of the `IB` and `OT` approaches yielding computationally favorable properties.

3.3 `IB+OT` — *Sinkhorn* Policy Learner

Suppose one were to discretize the `IB` problem, with objective in Equation 7, similarly to how `OT` was transformed into a linear program. The result might look like the finite-sample optimization problem detailed by Equation 9 with a mutual-information penalty added to the objective. Surprisingly, this is the exact problem that is efficiently solvable by the popular Sinkhorn algorithm (Cuturi, 2013; Peyré and Cuturi, 2019). Penalizing the mutual information is equivalent to rewarding the joint entropy whenever the marginals are fixed; hence, the information bottleneck corresponds to an entropic regularization

$$\begin{aligned} \hat{\pi}_\beta^{(n,m)} &= \arg\max_\pi \sum_{j=1}^{n}\sum_{i=1}^{m} (\hat{\varphi}_{ij}\pi_{ij} - \beta \log \pi_{ij}), \\ \text{s.t.} \quad & \frac{1}{m}\sum_{i=1}^{m} \pi_{ij} = 1, \quad \frac{1}{n}\sum_{j=1}^{n} \pi_{ij} = 1, \quad \pi_{ij} \geq 0. \end{aligned} \tag{11}$$

By duality theory (Boyd and Vandenberghe, 2004), for every $\beta \geq 0$ there exists some $\alpha \geq 0$ such that the admissible set of solutions is $\{\pi : I_\pi(\xi; U) \leq \alpha\}$, with $I_\pi(\cdots)$ from Equation 6. For the remainder of this paper, we largely concern ourselves with the finite-sample `OT` solver and recommend `IB+OT` for larger data.

3.4 Generalizing Policy Prescriptions

The finite-sample policy learners shown above, namely `OT` of Equation 10 and `IB+OT` of Equation 11, give prescriptions on the units in the second dataset split. By repeated application of our method with round-robin splits, as will be discussed further in §4, one obtains prescriptions for the entire sample. A *generalizable* prescription rule (policy) can be obtained by a regression on the in-sample prescriptions:

$$h_{\hat{\pi}}(u) \triangleq \hat{\mathbb{E}}[\, \mathbb{E}_{\hat{\pi}}\, \xi \mid U = u]. \tag{12}$$

This estimation is "easy" in the sense that it lacks the covariate/domain shift that tends to make causal-effect estimation difficult. $h_{\hat{\pi}}(u)$ is tasked with predicting the expected policy within the observational domain. There is no extrapolation on (T, X) combinations.

4 A SCALABLE ALGORITHM

Algorithm 1: Finite-sample Nudge Policy Learner

Input : $\{(y_j, t_j, x_j, u_j)\}_{j=1}^{n} \sim^{\text{i.i.d}} (Y, T, X, U)$, policy regularization $\gamma \geq 0$ or $\beta \geq 0$
Output : policy predictor $h_{\hat{\pi}}(u)$, Equation 12

foreach *k-fold partition of the training set* **do**
  Learn nuisances $(\hat{\mu}, \hat{\eta}_\xi)$ using data outside the current partition;
  Calibrate temperature of propensity-ratio $\hat{\eta}_\xi$ on the current partition;
Pool the corresponding nuisances across the partitions covering the entire training set;
Randomly re-partition the dataset into *policy batches* of equal cardinality;
foreach *policy batch* **do**
  Sample nudges $\{\xi_i\}_{i=1}^{m}$ from prior and compute pseudo-outcomes $\hat{\varphi}_{\xi_i}$, Equation 1;
  Optimize policy prescriptions $\hat{\pi}$ by solving linear program or Sinkhorn problem (Equations 10 / 11);
Estimate policy rule $h_{\hat{\pi}}(u)$ by regressing expected policy $\mathbb{E}_{\hat{\pi}}\, \xi$ on U on entire training set;

It is usually statistically favorable to partition the dataset into many small non-overlapping sets that each act as the second split (for pseudo-outcomes) in a problem instance with the rest of the data reserved for that problem's first split (for nuisances) (Chernozhukov et al., 2018). The benefit of this strategy is that it

avails more data for estimating the nuisances, and the drawback is the additional computation cost. For our method, estimating nuisances and pseudo-outcomes (§2) is only the first learning problem. We also seek to learn policies (§3), and it could be difficult to do so on small second splits. Therefore, we decouple these two data-partitioning design choices in Algorithm 1. For effect estimation, nuisances for pseudo-outcomes are estimated and then pooled across the whole sample. Then the policy is optimized in randomly allocated batches. Those prescriptions are pooled back together in order to generalize the policy out of sample according to the regression posed in Equation 12.

Whether the Sinkhorn or linear-program solver is used, it would be more efficient for large datasets to learn the policy in batches. Overall, for $\mathcal{O}(1)$ policy-batch sizes, Algorithm 1 achieves a runtime and memory complexity (in sample size) equivalent to that of the constituent machine-learning procedures.

5 ESTIMATION PROPERTIES

A recurrent quantity in our analysis shall be the conditional bias of the pseudo-outcome estimate for the heterogeneous nudge effect, $\hat{b}_\xi(X)$, defined in Lemma 1. This reveals the double (rate) robustness property.

Lemma 1 (Effect Estimation)**.** *The plug-in estimator for nudge effects $\hat{\varphi}_\xi$ shown in Equation 1 is locally doubly robust in the sense that only one of the two nuisances need be correctly specified; in other words, if either $\hat{\mu}(T,X) = \mu(T,X)$ or $\hat{\eta}_\xi(T,X) = \eta_\xi(T,X)$, then $\mathbb{E}[\hat{\varphi}_\xi \mid X] = \mathbb{E}[\Delta_\xi Y \mid X]$. Further, the estimator is* doubly rate robust *because for any $\{\hat{\mu}(T,X), \hat{\eta}_\xi(T,X)\}$,*

$$\hat{b}_\xi(X) \triangleq \mathbb{E}[\hat{\varphi}_\xi - \varphi_\xi \mid X] = \mathbb{E}[-(\hat{\mu} - \mu)(\hat{\eta}_\xi - \eta_\xi) \mid X].$$

In order to study the convergence of the learned policies, we introduce $\hat{\pi}^*$ that is the finite-sample solution with *oracle pseudo-outcomes* φ_ξ in place of their estimates $\hat{\varphi}_\xi$. This intermediate quantity shall allow us to separate the consequences of imperfect effect estimation from imperfect policy learning. Also, for finite-sample estimates like $\hat{\pi}$, the norm operator $\|\cdot\|_p$ is taken elementwise, like the Frobenius norm for $p = 2$, on the (m nudges $\times$ n units) matrices. For instance, $\|\hat{\pi}\|_2 = \sqrt{\sum_{i=1}^m \sum_{j=1}^n (\hat{\pi}_{ij})^2}$. The perturbation stability of linear programs (Robinson, 1980) lets us compare $\hat{\pi}$ and $\hat{\pi}^*$. We need one more assumption on uniqueness in order to proceed with the remaining lemmas. This assumption, while not guaranteed for our cost function in general, helps to simplify the theoretical results.

Assumption 2. The policies $\hat{\pi}$, $\hat{\pi}^\star$, $\pi^\star$ are uniquely optimal solutions to their respective programs.

Lemma 2. *For the vanilla `OT` policy-learning problem described by Equation 9, there exist positive constants (ε, δ) such that for any $\|\hat{\varphi} - \varphi\| < \delta$, one has*

$$\|\hat{\pi} - \hat{\pi}^*\| \leq \varepsilon \|\hat{\varphi} - \varphi\|.$$

In addition, $\mathbb{E}\,\|\hat{\varphi} - \varphi\|_1 \leq \sum_{i,j} \sqrt{(\hat{b}_{ij})^2 + (\hat{s}_{ij})^2}$ where $\hat{s}^2$ is the finite-sample conditional error variance,

$$\hat{s}_\xi(X)^2 \triangleq \mathrm{Var}[\hat{\varphi}_\xi - \varphi_\xi \mid X].$$

Finally, we study the asymptotics of the policy estimate through optimal transport (Villani et al., 2009).

Lemma 3. *Under mild conditions, transference plan $\hat{\pi}^*$ converges weakly to π^* as $n, m \to \infty$. If, in addition, $\hat{\eta}_\xi \to \eta_\xi$ and $\hat{\mu} \to \mu$ uniformly in $\mathcal{T} \times \mathcal{X}$ as $n \to \infty$, then $\hat{\pi}$ converges weakly to $\hat{\pi}^\star$ as well.*

Theorem 1 (Robust Policies)**.** *The policy-effect estimate is doubly rate robust in the sense that it can only be overestimated up to a product of the nuisance errors. In particular,*

$$\mathbb{E}[\mathbb{E}_{\hat{\pi}}\,\hat{\varphi}_\xi - \mathbb{E}_{\pi^*}\,\varphi_\xi] = \mathbb{E}[\mathbb{E}_{\hat{\pi}}\,\hat{b}_\xi] + \underbrace{\mathbb{E}[\mathbb{E}_{\hat{\pi}}\,\varphi_\xi - \mathbb{E}_{\hat{\pi}^*}\,\varphi_\xi]}_{\text{(A) } \textit{negative regret}} + \underbrace{\mathbb{E}[\mathbb{E}_{\hat{\pi}^*}\,\varphi_\xi - \mathbb{E}_{\pi^*}\,\varphi_\xi]}_{\text{(B) } \textit{discretization}} \quad (13)$$

with the first term, $\mathbb{E}[\mathbb{E}_{\hat{\pi}}\,\hat{b}_\xi]$ is the product of errors localized around the learned policy. Term (A) is nonpositive and term (B) is a discretization error unrelated to the effect estimation. Further, if the assumptions for Lemma 3 are satisfied,

$$\mathbb{E}[\mathbb{E}_{\hat{\pi}}\,\hat{\varphi}_\xi - \mathbb{E}_{\pi^*}\,\varphi_\xi] \to 0. \quad (14)$$

We show that the policy-effect estimate is consistent (Equation 14) and tends to be conservative (Equation 13), which is useful for prudent decision-making. Next, we shed light on how the nudge-marginal constraint discussed in §3 guarantees error localization.

Corollary 1.1 (Localized Errors)**.** *The error-product term $\mathbb{E}[\mathbb{E}_{\hat{\pi}}\,\hat{b}_\xi]$ in Theorem 1 can be understood in terms of the nudge prior $p(\xi)$, after observing the absolute bound*

$$\left|\mathbb{E}[\mathbb{E}_{\hat{\pi}}\,\hat{b}_\xi]\right| \leq \mathbb{E}_{p(\xi)} \sup_{u \in \mathcal{U}} \left|\hat{b}_\xi(u)\right|.$$

In words, the wort-case heterogeneous error is only relevant around the nudge prior.

Next, we seek intuition on the consequences of this bound via the nudge prior. Corollary 1.2 constructs an illustrative scenario with one nudge prior being more dispersed than another nudge prior.

Corollary 1.2 (Consequence of Nudge Dispersion)**.** *Suppose that $|\hat{b}_\xi(u)|$ is bounded above by some function $w(|\xi|, u)$ that is monotonically non-decreasing in*

every dimension of $|\xi|$ and has finite limit. Further, consider two candidate nudge priors p_1 and p_2 such that $\mathbb{P}_1[\|\xi\| > a] \leq \mathbb{P}_2[\|\xi\| > a]$ for every $a > 0$. Then the error bound of Corollary 1.1 cannot be greater for p_1 than for p_2.

6 EMPIRICAL EVALUATIONS

We sought to verify the empirical improvements conferred by the doubly robust estimation (§2) *and* the constrained policy learner (§3). Experiment source code may be found in the supplementary material.

6.1 Semi-synthetic Policy Learning

Learner	Policy Effect Mean	(Std.Err.)
Robust `OT` (Alg. 1)	**0.80**	(0.05)
Naïve `OT`	−0.01	(0.01)
Robust Full	−1.49	(0.16)
Naïve Full	−1.49	(0.16)

Table 1. Mean (and standard error) of policy effects on 20 trials (seeds 0–19) of the TCGA semi-synthetic benchmark. Our robust `OT` learner with $k = 5$ dataset splits is compared to three ablated baselines that do not achieve policy improvement. Units are in standard deviations of the observational outcomes.

The first benchmark we conducted was built off of the cancer genome atlas (TCGA) dataset originally proposed for causal-estimation benchmarking by Bica et al. (2020). Our semi-synthetic causal setting was designed to resemble a wide diversity of the problems mentioned in §1. Concretely, we aimed for real-valued multivariate treatments with significant covariate shift and complex response surfaces that reach a floor (of zero) at extreme-valued treatments. These are all reasonable traits for real-life policy learning from observational studies.

The baselines considered in Table 1 were ablations to Algorithm 1. The Robust/Naïve demarcation indicated whether $\hat{\varphi}_\xi$ was used or the direct prediction $\Delta_\xi \hat{\mu}$. On the other hand, `OT` → Full indicated a policy learned by directly optimizing the effect estimate rather than regression on the OT transference plan. Figure 2 shows the change in performance at different nudge priors.

To verify that the policy learner of Equation 12 abides by the nudge prior, we present Figure 3 comparing the quantiles of the empirical nudges with the prior. They appear to align overall, although the learned prescriptions are slightly under-dispersed, which is reasonable.

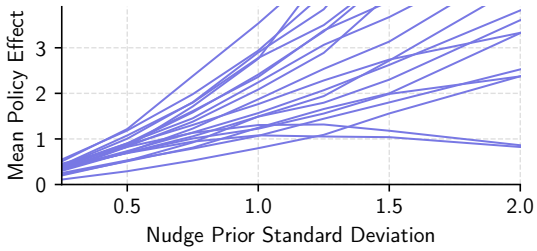

Figure 2. Achieved TCGA policy effects for different nudge priors (interventional budgets), across the 20 random seeds. Table 1 reports evaluations for nudge-prior standard deviation set to 0.5. We observe increasing policy-effect variance with increasing nudge dispersion.

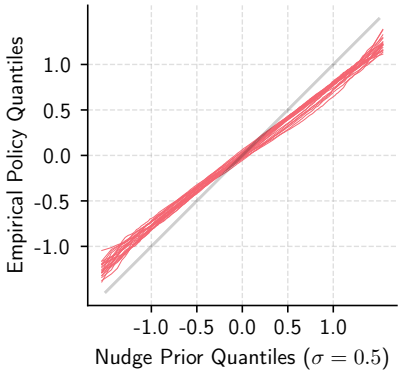

Figure 3. A quantile-quantile plot of the TCGA nudge prior compared with the empirical distribution of the 20 learned policy prescriptions applied out of sample.

6.2 Experiment with Yelp & IRS Data

The second benchmark used latitudes and longitudes of Yelp establishments across the United States as the "treatment" variable (Yelp, 2023). In this case, the treatment assignment is *not* synthetic and indeed rather complex because it relies on the geography of major cities in the country. The outcome was a semi-synthetic imitation of revenue that depended on features like Yelp reviews and income brackets of the establishment's ZIP Code (Internal Revenue Service, 2020). Here, a nudge effect is defined as the change in revenue from moving location, using a Gaussian nudge prior with 1° of standard deviation.

We trained nuisance models on half the dataset and learned policies on the other half in order to compare variations of the effect estimator. All policies were learned by `OT` with $\gamma = 0$, so the experiment in Table 2 served to contrast the proposed estimator against the naïve direct estimate, and the robust estimate with an SMD-trained (Equation 3) propensity in place of the recommended PC of Equation 2.

Learner	Policy Effect Mean	(Std.Err.)
Robust `OT`	**0.39**	(0.03)
Naïve `OT`	0.27	(0.002)
SMD `OT`	0.27	(0.002)

Table 2. Mean (and standard error) of policy effects on 10 trials of the Yelp & IRS semi-synthetic benchmark. Both the robust estimator equipped with an alternative SMD-trained propensity, and the naïve estimator, do not improve upon the proposed robust estimator.

7 DISCUSSION

Our analyses focus on the `OT` (§3.2) policy-learning scheme, although we began §3 by building `IB` (§3.1) in order to justify `IB+OT` (§3.3) down the line. Our results validate `OT` and, moving forward, suggest the general utility of the Sinkhorn algorithm for solving the entropically regularized version, which is `IB+OT`.

In our analysis of the estimation properties (§5), we revealed fundamental properties of the learned policy in relation to oracle pseudo-outcomes. Theorem 1 deconstructs the policy-effect error into understandable components, and its corollaries explore basic phenomena arising from the choice of nudge prior.

In our empirical evaluations (§6), we demonstrated the necessity of each ingredient for the proposed approach. The TCGA benchmark (§6.1) showed that a learned prescription rule applied out of sample requires the added robustness of pseudo-outcomes along with the `OT` policy-learning scheme to reliably achieve positive policy effects. The Yelp & IRS experiment (§6.2) that the propensity nuisance for the pseudo-outcome estimate must be learned using the novel PC loss of Equation 2 for significantly higher policy effects.

Possible extensions. Returning to the wildfire example laid out in §1, we are struck with the need to deal with two difficulties often faced in real-life causal inference: treatment interference, and hidden confounding. Spatially proximal observation units, like patches of land, can be affected by one another's interventions. Also, not all confounders are recorded—much less perfectly. Both of these issues can be handled with extensions to the proposed nudge-policy framework.

Future work. We recognize the need to further study the consequences of using pseudo-outcomes as an approximation of the truly sought optimal-transport objective. We plan to augment the proposed policy learners in a way that gives stronger statistical guarantees on the actual policy effect. This would likely entail alterations to the optimization problem in order to include a conditional independence constraint, ensuring that the transference plan is only a function of the relevant covariates U, and not T.

8 CONCLUSION

We present a set of algorithms for learning continuous and multivariate policies for *localized* interventions, from offline observational data. We demonstrated our Algorithm 1 on semi-synthetic datasets against baselines that showcase the benefits offered by each component of the algorithm. Further, our Theorem 1 supplies an interpretable guarantee on the policy-effect error localized by the nudge prior, which can be dictated by interventional budgets or other domain knowledge.

Acknowledgments

This work was funded in part by Defense Advanced Research Projects Agency (DARPA) and Army Research Office (ARO) under Contract No. W911NF-21-C-0002.

References

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

I. Bica, J. Jordon, and M. van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33:16434–16445, 2020.

S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018.

V. Chernozhukov, W. K. Newey, and R. Singh. Debiased machine learning of global and local parameters using regularized riesz representers. *The Econometrics Journal*, 25(3):576–601, 2022.

K. Choi, C. Meng, Y. Song, and S. Ermon. Density ratio estimation via infinitesimal classification. In *International Conference on Artificial Intelligence and Statistics*, pages 2552–2573. PMLR, 2022.

K. Colangelo and Y.-Y. Lee. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*, 2020.

F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

G. Dantzig. *Linear programming and extensions*. Princeton university press, 1963.

M. Demirer, V. Syrgkanis, G. Lewis, and V. Chernozhukov. Semi-parametric efficient policy learning with continuous actions. *Advances in Neural Information Processing Systems*, 32, 2019.

M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012.

B. Ewert. Moving beyond the obsession with nudging individual behaviour: Towards a broader understanding of behavioural public policy. *Public Policy and Administration*, 35(3):337–360, 2020.

M. Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*, 2016.

O. Hines, O. Dukes, K. Diaz-Ordaz, and S. Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76(3):292–304, 2022.

O. Hines, K. Diaz-Ordaz, and S. Vansteelandt. Optimally weighted average derivative effects. *arXiv preprint arXiv:2308.05456*, 2023.

A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4), 2005.

H. Ichimura and W. K. Newey. The influence function of semiparametric estimators. *Quantitative Economics*, 13(1):29–61, 2022.

Internal Revenue Service. 2020 zip code data (soi), 2020. URL https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2020-zip-code-data-soi.

N. Kallus and A. Zhou. Policy evaluation and optimization with continuous treatments. In *International conference on artificial intelligence and statistics*, pages 1243–1251. PMLR, 2018.

N. Kallus, X. Mao, K. Wang, and Z. Zhou. Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning*, pages 10598–10632. PMLR, 2022.

L. V. Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

E. H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

E. H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*, 2022.

M. Liang, T. Ye, and H. Fu. Estimating individualized optimal combination therapies through outcome weighted deep learning algorithms. *Statistics in medicine*, 37(27):3869–3886, 2018.

M. G. Marmarelis, E. Haddad, A. Jesson, N. Jahanshad, A. Galstyan, and G. Ver Steeg. Partial identification of dose responses with hidden confounders. In *Uncertainty in Artificial Intelligence*, pages 1368–1379. PMLR, 2023.

L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

L. Nie, M. Ye, qiang liu, and D. Nicolae. {VCN}et and functional targeted regularization for learning causal effects of continuous treatments. In *International Conference on Learning Representations*, 2021.

M. Oprescu, J. Dorn, M. Ghoummaid, A. Jesson, N. Kallus, and U. Shalit. B-learner: Quasi-oracle bounds on heterogeneous causal effects under hidden confounding. *arXiv preprint arXiv:2304.10577*, 2023.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

G. Peyré and M. Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

S. M. Robinson. Strongly regular generalized equations. *Mathematics of Operations Research*, 5(1): 43–62, 1980.

K. B. Rodgers, S.-S. Lee, N. Rosenbloom, A. Timmermann, G. Danabasoglu, C. Deser, J. Edwards,

J.-E. Kim, I. R. Simpson, K. Stein, M. F. Stuecker, R. Yamaguchi, T. Bódai, E.-S. Chung, L. Huang, W. M. Kim, J.-F. Lamarque, D. L. Lombardozzi, W. R. Wieder, and S. G. Yeager. Ubiquity of human-induced changes in climate variability. *Earth System Dynamics*, 12(4):1393–1411, 2021. doi: 10.5194/esd-12-1393-2021.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

K. Swersky, M. Ranzato, D. Buchman, N. D. Freitas, and B. M. Marlin. On autoencoders and score matching for energy based models. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1201–1208, 2011.

R. J. Tibshirani, R. Foygel Barber, E. Candes, and A. Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

M. Uehara, M. Kato, and S. Yasui. Off-policy evaluation and learning for external validity under a covariate shift. *Advances in Neural Information Processing Systems*, 33:49–61, 2020.

C. Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

P. Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7): 1661–1674, 2011.

Q. Xu, X. Cao, G. Chen, H. Zeng, H. Fu, and A. Qu. Multi-label residual weighted learning for individualized combination treatment rule. *arXiv preprint arXiv:2310.00864*, 2023a.

Q. Xu, H. Fu, and A. Qu. Optimal individualized treatment rule for combination treatments under budget constraints. *arXiv preprint arXiv:2303.11507*, 2023b.

Z. Yang, Y. Guo, P. Xu, A. Liu, and A. Anandkumar. Distributionally robust policy gradient for offline contextual bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 6443–6462. PMLR, 2023.

Yelp. Yelp open dataset, 2023. URL https://www.yelp.com/dataset.

Z. Zhou, S. Athey, and S. Wager. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. *Yes: problem setting and assumptions mainly in §2, algorithm in §4, and nuisance model parametrizations in §B.*
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. *Yes, in §4.*
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. *Yes, provided as a .zip file in the supplementary material.*

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. *Yes, largely in §5 besides the basic assumptions in §2.*
   (b) Complete proofs of all theoretical results. *Yes, in §A of the appendix.*
   (c) Clear explanations of any assumptions. *Yes, mostly pertaining to the problem setup described in §2 and other mild conditions in §5.*

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). *Yes, in the supplementary material.*
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). *Yes, in §B of the appendix.*
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). *Yes, random seeds are listed in §6.*
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). *Yes, in §B.*

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator if your work uses existing assets. *Yes, citing those who previously released the datasets that we consider in §6.*

(b) The license information of the assets, if applicable. *Yes, in particular the Yelp license described in §B.*

(c) New assets either in the supplemental material or as a URL, if applicable. *Yes, source code included in supplementary material.*

(d) Information about consent from data providers/curators. *Yes: link to terms of use for Yelp dataset included in §B.*

(e) Discussion of sensitive content if applicable, e.g., personally identifiable information or offensive content. *Not Applicable*

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. *Not Applicable*

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. *Not Applicable*

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. *Not Applicable*

A PROOFS

A.1 Proof of Proposition 1

Following closely the approaches of Ichimura and Newey (2022) and Hines et al. (2022), we give a brief exposition on deriving the efficient influence function for $\mathbb{E}[\Delta_\xi Y]$. The notation for these semiparametric estimators is a little different than in our paper. For instance, we denote the oracle estimator as $\mathcal{P}$ and the oracle estimate for a parameter θ as $\theta(\mathcal{P})$. We compare these quantities to those from arbitrary estimators $\tilde{\mathcal{P}}$ and the empirical plug-in estimator $\hat{\mathcal{P}}_n$. With nudge effects, we have $\theta(\mathcal{P}) \triangleq \mathbb{E}[\Delta_\xi Y]$ for some ξ.

$$\theta(\mathcal{P}) = \mathbb{E}[\,\mathbb{E}[Y \mid T = T + \xi,\ X] - \mathbb{E}[Y \mid T,\ X]\,],$$
$$\theta(\hat{\mathcal{P}}_n) = n^{-1} \sum_{j=1}^{n} \Big(\hat{\mathbb{E}}_n[Y \mid T = t_j + \xi,\ X = x_j] - \hat{\mathbb{E}}_n[Y \mid T = t_j,\ X = x_j] \Big).$$

Then we look at a *parametric submodel* with some perturbation to the oracle estimator: $\mathcal{P}_\tau = \tau\tilde{\mathcal{P}} + (1-\tau)\mathcal{P}$ for $\tau \in [0,1]$. Let us assume that $\tilde{\mathcal{P}}$ is a point mass in some $\tilde{w} = (\tilde{y}, \tilde{t}, \tilde{x})$. We calculate the following Gateaux derivative:

$$\left.\frac{\mathrm{d}\theta(\mathcal{P}_\tau)}{\mathrm{d}\tau}\right|_{\tau=0} \triangleq \psi(\tilde{w}, \mathcal{P}) \quad \text{termed canonical gradient at } \tilde{w}\text{, through Riesz representer theorem.}$$

Now denoting the density estimates using $\mathcal{P}_\tau$ as $f_\tau(\dots)$, it holds under mild regularity conditions that

$$\begin{aligned}\theta(\mathcal{P}_\tau) &= \iiint \Big[y f_\tau(y|t+\xi, x) f_\tau(t,x) - y f_\tau(y|t,x) f_\tau(t,x) \Big] \,\mathrm{d}y\,\mathrm{d}t\,\mathrm{d}x, \\ &= \iiint \Big[f_\tau(y|t+\xi,x) - f_\tau(y|t,x) \Big] y f_\tau(t,x)\,\mathrm{d}y\,\mathrm{d}t\,\mathrm{d}x.\end{aligned}$$

To zoom in on one of these terms, we denote $\theta_\xi(\mathcal{P}_\tau) = \iiint f_\tau(y|t+\xi,x)\, y f_\tau(t,x)\,\mathrm{d}y\,\mathrm{d}t\,\mathrm{d}x$. Now, taking the derivative,

$$\begin{aligned}\left.\frac{\mathrm{d}\theta_\xi(\mathcal{P}_\tau)}{\mathrm{d}\tau}\right|_{\tau=0} = \iiint \Bigg\{ &\frac{f_{\tau=0}(t,x)}{f_{\tau=0}(t+\xi,x)} \frac{\mathrm{d}}{\mathrm{d}\tau} f_\tau(y, t+\xi, x)|_{\tau=0} \\ &- \frac{f_{\tau=0}(t,x) f_{\tau=0}(y,t+\xi,x)}{f_{\tau=0}(t+\xi,x)^2} \frac{\mathrm{d}}{\mathrm{d}\tau} f_\tau(t+\xi,x)|_{\tau=0} \\ &+ \frac{f_{\tau=0}(y,t+\xi,x)}{f_{\tau=0}(t+\xi,x)} \frac{\mathrm{d}}{\mathrm{d}\tau} f_\tau(t,x)|_{\tau=0} \Bigg\} y\,\mathrm{d}y\,\mathrm{d}t\,\mathrm{d}x\end{aligned}$$

By using the identity $\mathrm{d}f_\tau(w)/\,\mathrm{d}\tau|_{\tau=0} = \delta_{\tilde{w}}(w) - f_{\tau=0}(w)$ (Hines et al., 2022), which means for nudges that

$$\left.\frac{\mathrm{d}f_\tau(t+\xi,x)}{\mathrm{d}\tau}\right|_{\tau=0} = \delta_{(\tilde{t},\tilde{x})}(t+\xi,x) - f_{\tau=0}(t+\xi,x),$$

$$\begin{aligned}\left.\frac{\mathrm{d}\theta_\xi(\mathcal{P}_\tau)}{\mathrm{d}\tau}\right|_{\tau=0} &= \iiint y f_{\tau=0}(y|t+\xi,x) f_{\tau=0}(t,x) \left\{ \frac{\delta_{\tilde{w}}(y,t+\xi,x)}{f_{\tau=0}(y,t+\xi,x)} - \frac{\delta_{(\tilde{t},\tilde{x})}(t+\xi,x)}{f_{\tau=0}(t+\xi,x)} + \frac{\delta_{(\tilde{t},\tilde{x})}(t,x)}{f_{\tau=0}(t,x)} - 1 \right\}, \\ &= \frac{f_{\tau=0}(\tilde{t}-\xi,\tilde{x})}{f_{\tau=0}(\tilde{t},\tilde{x})}\tilde{y} - \frac{f_{\tau=0}(\tilde{t}-\xi,\tilde{x})}{f_{\tau=0}(\tilde{t},\tilde{x})}\mathbb{E}[Y \mid T = \tilde{t},\ X = \tilde{x}] + \mathbb{E}[Y \mid T = \tilde{t}+\xi,\ X = \tilde{x}] - \theta_\xi(\mathcal{P}), \\ &= \eta_\xi(\tilde{t},\tilde{x})\Big[\tilde{y} - \mu(\tilde{t},\tilde{x})\Big] + \mu(\tilde{t}+\xi,\tilde{x}) - \theta_\xi(\mathcal{P}). \\ \therefore \left.\frac{\mathrm{d}\theta(\mathcal{P}_\tau)}{\mathrm{d}\tau}\right|_{\tau=0} &= \Big[\eta_\xi(\tilde{t},\tilde{x}) - 1\Big]\Big[\tilde{y} - \mu(\tilde{t},\tilde{x})\Big] + \Delta_\xi\mu(\tilde{t},\tilde{x}) - \theta_\xi(\mathcal{P}).\end{aligned}$$

This result proves the proposition. The way the efficient influence function is used to form a corrective term in a doubly robust estimate is by the following von Mises expansion,

$$\theta(\mathcal{P}) = \theta(\tilde{\mathcal{P}}) - \left.\frac{\mathrm{d}\theta(\mathcal{P}_\tau)}{\mathrm{d}\tau}\right|_{\tau=1} + \text{remainder}.$$

Notice that the Gateaux derivative is evaluated at $\tau = 1$ rather than $\tau = 0$. It is known that

$$\left.\frac{\mathrm{d}\theta(\mathcal{P}_\tau)}{\mathrm{d}\tau}\right|_{\tau=1} = -\mathcal{P}\{\psi(W, \tilde{\mathcal{P}})\},$$

which we estimate as $\mathcal{P}\{\psi(W, \hat{\mathcal{P}}_n)\} \approx n^{-1}\sum_{j=1}^n \psi(w_j, \hat{\mathcal{P}}_n)$ when $\tilde{P} = \hat{\mathcal{P}}_n$. Finally, $\theta(\mathcal{P}) \approx \hat{\varphi}_\xi$ harkening back to Equation 1, since $\hat{\varphi}_\xi = \hat{\mathcal{P}}_n\{\psi(W, \hat{\mathcal{P}}_n)\} + \theta(\hat{\mathcal{P}}_n)$.

A.2 Proof of Lemma 1

These proofs rely largely on iterated expectations. We first show that the conditional expectation of the pseudo-outcome $\hat{\varphi}_\xi$ is doubly robust to nuisance misspecification. There are two branches in this proof, from either considering (a) $\hat{\mu} = \mu$ or (b) $\hat{\eta}_\xi = \eta_\xi$.

For $(\hat{\mu} = \mu)$, the expectation is quite simple: *(arguments T, X to the parameters omitted for brevity)*

$$\begin{aligned}\mathbb{E}[\hat{\varphi}_\xi \mid X] = \mathbb{E}[(\hat{\eta}_\xi - 1)(Y - \mu) + \Delta_\xi\mu \mid X] &= \mathbb{E}_T\big[\,\mathbb{E}_Y[(\hat{\eta}_\xi - 1)(Y - \mu) \mid T, X] + \Delta_\xi\mu \mid X\big],\\ &= \mathbb{E}_T[(\hat{\eta}_\xi - 1)(\mu - \mu) + \Delta_\xi\mu \mid X] = \mathbb{E}[\Delta_\xi Y \mid X].\end{aligned}$$

The $(\hat{\eta}_\xi = \eta_\xi)$ branch is slightly more involved,

$$\begin{aligned}\mathbb{E}[(\eta_\xi - 1)(Y - \hat{\mu}) + \Delta_\xi\hat{\mu} \mid X] &= \mathbb{E}_T[(\eta_\xi - 1)(\mu - \hat{\mu}) + \Delta_\xi\hat{\mu} \mid X],\\ &= \mathbb{E}_T\Big[\eta_\xi(T, X)\big[\mu(T, X) - \hat{\mu}(T, X)\big] + \hat{\mu}(T + \xi, X) - \mu(T, X) \;\Big|\; X\Big].\end{aligned}$$

Lemma 0. Notice that for any integrable $v(T, X)$, we have

$$\begin{aligned}\mathbb{E}[\eta_\xi v \mid X] &= \int_{\mathcal{T}} \eta_\xi(t, X)v(t, X)f_0(t|X)\,\mathrm{d}t,\\ &= \int_{\mathcal{T}} \frac{f_0(t - \xi|X)}{f_0(t|X)}v(t, X)f_0(t|X)\,\mathrm{d}t,\\ &= \int_{\mathcal{T}} f_0(t'|X)\,v(t' + \xi, X)\,\mathrm{d}t' \qquad \text{where } t' = t - \xi,\\ &= \mathbb{E}[v(T + \xi, X) \mid X]. \quad \square\end{aligned}$$

We have leveraged the assumption that $\mathcal{T}$ is unbounded.

Hence,

$$\begin{aligned}&\mathbb{E}_T\big[\eta_\xi(T, X)[\mu(T, X) - \hat{\mu}(T, X)] \;\big|\; X\big] = \mathbb{E}_T\big[\mu(T + \xi, X) - \hat{\mu}(T + \xi, X) \;\big|\; X\big],\\ \therefore\quad &\mathbb{E}[(\eta_\xi - 1)(Y - \hat{\mu}) + \Delta_\xi\hat{\mu} \mid X] = \mathbb{E}_T\big[\mu(T + \xi, X) - \mu(T, X) \;\big|\; X\big] = \mathbb{E}[\Delta_\xi Y \mid X].\end{aligned}$$

Double rate robustness. We follow a similar path for the double rate robustness property, revealed as a product of errors in $\hat{b}_\xi(X) \triangleq \mathbb{E}[\hat{\varphi}_\xi - \varphi_\xi \mid X]$. To start, we examine the difference in the corrective term of the pseudo-outcome, $(\eta_\xi - 1)(Y - \mu)$, coincidentally the only part that depends on Y:

$$\mathbb{E}_Y\big[(\hat{\eta}_\xi - 1)(Y - \hat{\mu}) - (\eta_\xi - 1)(Y - \mu) \mid T, X\big] = (\hat{\eta}_\xi - 1)(\mu - \hat{\mu}).$$

Again, when arguments are omitted, (T, X) are implied. The fact above allows us to short-circuit to

$$\begin{aligned}\hat{b}_\xi(X) = \mathbb{E}_T\big[(\hat{\eta}_\xi - 1)(\mu - \hat{\mu}) + \Delta_\xi\hat{\mu} - \Delta_\xi\mu \mid X\big] &= \mathbb{E}_T\big[\hat{\eta}_\xi \times (\mu - \hat{\mu}) + \hat{\mu}(T + \xi, X) - \mu(T + \xi, X) \mid X\big],\\ &= \mathbb{E}_T\big[\hat{\eta}_\xi \times (\mu - \hat{\mu}) + \eta_\xi\hat{\mu} - \eta_\xi\mu \mid X\big] \quad \textit{by Lemma 0 above,}\\ &= \mathbb{E}_T\big[-(\hat{\eta}_\xi - \eta_\xi)(\hat{\mu} - \mu) \mid X\big].\end{aligned}$$

A.3 Proof of Lemma 2

This result follows from the main theorem of Robinson (1980). By construction, what distinguishes the finite-sample $\hat{\pi}$ from its oracle intermediary $\hat{\pi}^*$ is the substitution $\hat{\varphi}_\xi \mapsto \varphi_\xi$. As long as the solution set for the linear program posed in Equation 9 is nonempty and bounded, then the (ε, δ) condition asserted in the lemma holds.

The second part of the lemma, decomposing the absolute norm in terms of the conditional bias $\hat{b}_\xi$ and the conditional error variance $\hat{s}_\xi^2$, follows from a simple norm inequality.

By the definition of variance, we have $\mathbb{E}[(\hat{\varphi}_{ij} - \varphi_{ij})^2] = (\hat{b}_{ij})^2 + (\hat{s}_{ij})^2$. Recall that φ_{ij} is shorthand for $\varphi_{\xi_i}(u_j)$, and so on. Additionally, by Jensen's inequality, $\mathbb{E}\,|\hat{\varphi}_{ij} - \varphi_{ij}| \leq \sqrt{\mathbb{E}(\hat{\varphi}_{ij} - \varphi_{ij})^2}$. Therefore,

$$\mathbb{E}\,|\hat{\varphi}_{ij} - \varphi_{ij}| \leq \sqrt{(\hat{b}_{ij})^2 + (\hat{s}_{ij})^2} \quad \iff \quad \mathbb{E}\,\|\hat{\varphi} - \varphi\|_1 \leq \sum_{i=1}^{m}\sum_{j=1}^{n} \sqrt{(\hat{b}_{ij})^2 + (\hat{s}_{ij})^2}.$$

A.4 Proof of Lemma 3

This result relies on Theorem 5.20 of Villani et al. (2009). To satisfy all the conditions of convergence, we must assume that the finite-sample and infinite-sample oracle transference plans $\hat{\pi}^*$ and π^* are both uniquely optimal for their respective problems.

First we investigate the requirements for $\hat{\pi}^* \to \pi^*$ weakly. The (oracle) cost functions are identical here: $c(\xi, W) = -\varphi_\xi(W)$. However, one is finite-sample and the other is infinite. The stability theorem of optimal transport requires that the marginals converge weakly. In our case, the empirical nudge sample and the empirical observational sample both converge weakly to their oracle equivalents.

Next we prove that $\hat{\pi} \to \hat{\pi}^*$. The corresponding marginals for $\hat{\pi}$ and $\hat{\pi}^*$ are identical, simplifying the conditions for convergence. Since we presupposed uniform convergence of the nuisance parameters, we have uniform convergence in the pseudo-outcome $\hat{\varphi}_\xi \to \varphi_\xi$ as well. Hence the cost function converges uniformly and all the conditions for transference-plan convergence are satisfied.

A.5 Proof of Theorem 1

The decomposition of Equation 13 is straightforward to derive. The key ingredient is showing that, for any π,

$$\mathbb{E}[\mathbb{E}_\pi\, \hat{\varphi}_\xi] = \mathbb{E}[\mathbb{E}_\pi\, \varphi_\xi] + \mathbb{E}[\mathbb{E}_\pi(\hat{\varphi}_\xi - \varphi_\xi)] = \mathbb{E}[\mathbb{E}_\pi\, \varphi_\xi] + \mathbb{E}[\mathbb{E}_\pi\, \hat{b}_\xi].$$

Then, by setting $\pi = \hat{\pi}$, we obtain $\mathbb{E}[\mathbb{E}_{\hat{\pi}}\, \hat{\varphi}_\xi - \mathbb{E}_{\pi^*}\, \varphi_\xi] = \mathbb{E}[\mathbb{E}_\pi\, \hat{b}_\xi] + \mathbb{E}[\mathbb{E}_{\hat{\pi}}\, \varphi_\xi - \mathbb{E}_{\pi^*}\, \varphi_\xi]$, and the rest follows from linearity of expectation. Equipped additionally with Lemma 3, we can show that every one of these expectations vanishes asymptotically as in Equation 14. We restate the main decomposition;

$$\mathbb{E}[\mathbb{E}_{\hat{\pi}}\, \hat{\varphi}_\xi - \mathbb{E}_{\pi^*}\, \varphi_\xi] = \mathbb{E}[\mathbb{E}_{\hat{\pi}}\, \hat{b}_\xi] + \underbrace{\mathbb{E}[\mathbb{E}_{\hat{\pi}}\, \varphi_\xi - \mathbb{E}_{\hat{\pi}^*}\, \varphi_\xi]}_{\text{(A) negative regret}} + \underbrace{\mathbb{E}[\mathbb{E}_{\hat{\pi}^*}\, \varphi_\xi - \mathbb{E}_{\pi^*}\, \varphi_\xi]}_{\text{(B) discretization}}$$

By uniform convergence in the nuisances, we have the first error-localization term converge to zero. Both (A) and (B) likewise converge due to the weak respective convergences in the transference plans as stated by Lemma 3.

A.6 Proof of Corollary 1.1

Firstly we notice that $\mathbb{E}[\mathbb{E}_{\hat{\pi}}\, \hat{b}_\xi] = \int_{\mathcal{T}} \mathbb{E}[\hat{b}_\xi(U)\hat{\pi}(\xi|U)]\, \mathrm{d}\xi$. Taking absolute values,

$$\begin{aligned}
\left|\mathbb{E}[\mathbb{E}_{\hat{\pi}}\, \hat{b}_\xi]\right| \leq \int_{\mathcal{T}} \mathbb{E}\left|\hat{b}_\xi(U)\hat{\pi}(\xi|U)\right| \mathrm{d}\xi &= \int_{\mathcal{T}} \left\|\hat{b}_\xi \hat{\pi}\right\|_1 \mathrm{d}\xi && \text{(where this norm is over the expectation)} \\
&\leq \int_{\mathcal{T}} \left\|\hat{b}_\xi\right\|_\infty \|\hat{\pi}\|_1 \,\mathrm{d}\xi && \text{(by Hölder's inequality)} \\
&= \int_{\mathcal{T}} \sup_{u \in \mathcal{U}} \left\{|\hat{b}_\xi(u)|\right\} p(\xi)\, \mathrm{d}\xi && \text{(by the prior constraint on the marginal)} \\
&= \mathbb{E}_{p(\xi)} \sup_{u \in \mathcal{U}} \left|\hat{b}_\xi(u)\right|.
\end{aligned}$$

A.7 Proof of Corollary 1.2

We begin by proving the univariate-treatment case, for simplicity. The univarite nudge will be denoted as ξ^0.

Since $w(|\xi^0|, u)$ is monotonically non-decreasing in $|\xi^0|$, so is the function $v(\xi^0) \triangleq \sup_{u\in\mathcal{U}} w(|\xi^0|, u)$, which is an upper bound of $\xi^0 \mapsto \sup_{u\in\mathcal{U}} |\hat{b}_{\xi^0}(u)|$. For nudge-prior CDFs (F_1, F_2) we have $F_1(\xi^0) \geq F_2(\xi^0)$ for all $\xi^0 > 0$ and $F_1(\xi^0) \leq F_2(\xi^0)$ for all $\xi^0 < 0$, by the dispersion assumption. Also, for $k \in \{1, 2\}$,

$$\mathbb{E}_{p_k} v(\xi^0) = \int_{-\infty}^{+\infty} v(\xi^0) p_k(\xi^0)\,\mathrm{d}\xi^0 = \Big[v(\xi^0)F_k(\xi^0)\Big]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} \dot{v}(\xi^0)F_k(\xi^0)\,\mathrm{d}\xi^0$$

$$\begin{aligned}\therefore \mathbb{E}_{p_2} v(\xi^0) - \mathbb{E}_{p_1} v(\xi^0) &= \underbrace{[v(+\infty) - v(+\infty)]}_{\to 0} + \int_{-\infty}^{+\infty} \dot{v}[F_1 - F_2]\,\mathrm{d}\xi^0 \qquad \text{(by finite limit assumption)} \\ &= \underbrace{\int_0^{+\infty} \dot{v}[F_1 - F_2]\,\mathrm{d}\xi^0}_{\geq 0} - \underbrace{\int_0^{-\infty} \dot{v}[F_1 - F_2]\,\mathrm{d}\xi^0}_{\leq 0} \geq 0.\end{aligned}$$

To generalize to multivariate nudges, we bring in the multivariate densities $p_1(\xi), p_2(\xi)$ as well as their multivariate CDFs. The above inequality can be attained for vectors ξ by iterated integration on each dimension.

B EXPERIMENTAL SETUP

All experiments were performed in an internal cluster of Intel Xeon servers with Nvidia 1080Ti GPUs.

B.1 TCGA

There were 9,659 individuals in this panel dataset with sampled expressions for 4,000 genes. Like in previous usages of TCGA for causal-effect benchmarking, we projected the genes into a smaller set of variables. In our case we built the outcomes as random polynomials in these random projections. Namely, with all Z variables denoting matrices with i.i.d standard Gaussian random entries, and G the (genes $\times$ units) expression matrix, we projected

$$\tilde{X} = Z^x G, \quad \tilde{T} = (Z^t + Z^{x\to t} Z^x) G$$

and let X and T be the z-score normalized versions of these projections. We had 60 covariates and 4 treatments for the purpose of our experiments. It follows that, $Z^x \in \mathbb{R}^{60\times 4000}$, $Z^t \in \mathbb{R}^{4\times 4000}$, and $Z^{x\to t} \in \mathbb{R}^{4\times 60}$. The latter is a mixing matrix that increases dependence between X and T, i.e. covariate shift. Now, by stretching the notation, we also considered $Z^{x\to y} \in \mathbb{R}^{4\times 60}$ and $Z^{t\to y} \in \mathbb{R}^{4\times 4}$ to project the observed variables into 4 latent variables that passed through nonlinearities to form the outcome, which took the structural form

$$Y = \text{softplus}\left(-\sum_{b=1}^{4}\Big(\frac{Z^{x\to y} X\sqrt{4/60} + Z^{t\to y} T\sqrt{60/4}}{\sqrt{64}}\Big)^b - \|[X\ T]\|_2^6 + S\right), \quad \text{where } S \sim \text{Normal}(0, 0.1^2).$$

The scaling terms serve to weigh the treatment and covariates similarly. The norm on the concatenated covariates and treatments exists to ensure that extreme treatments push the outcome towards zero and not $+\infty$.

Estimation. As for the specific invocation of Algorithm 1, we estimated nuisances on a classic 5-fold dataset split using typical 2-layer, 50-unit, SiLU-activated feedforward neural networks (as well as for the final step of policy regression.) Batches had size 256, training always ran for 1024 epochs, nudge sample sizes were also 1024 for the `OT` step, and ADAM learning rates were set to 1×10^{-5} for the outcome model and 5×10^{-5} for everything else by simple hyperparameter search via the validation sets. When `OT` was used, we employed the rendition of Equation 10 with $\gamma = 0$ (for no L^1 regularization). This way, we did not have to tune any further hyperparameters, and we could efficiently repeat the whole experimental setup for many random seeds.

B.2 Yelp & IRS

There were 150,243 restaurants with valid entries in the Yelp dataset (Yelp, 2023). We extracted five-dimensional UMAP (McInnes et al., 2018) embeddings from BERTopic (Grootendorst, 2022) on the establishments' category fields. These counted as covariates. They were supplemented by the establishments' number of reviews, average score, and whether they were open or permanently closed. For the purpose of this section, we denote all those covariates as $X_{1:5}$ (UMAP), X_6, X_7, X_8 respectively. The treatment, T, is (latitude, longitude) of the establishment. We also obtained resident income information for the ZIP codes of T via an IRS dataset (Internal Revenue Service, 2020). The number of people (field "N2") in the top two brackets, #6 and #5, were used in constructing the link to the synthetic revenue outcome variable. We denote these count variables as $R_1(T), R_2(T)$.

The revenue outcome is modeled as

$$Y = \frac{[10R_1(T) + X_8 R_2(T)]\left(\sqrt{X_6} + S\right) X_7}{\|X_{1:5}\|_2}, \quad \text{where } S \sim \text{Exp}(\mu = 10),$$

which heterogeneously relates revenue to neighborhood income and Yelp reviews, incorporating nonlinear noise S.

Nuisance estimation. We used XGBoost (Chen and Guestrin, 2016) with 5-fold cross-validation to select an outcome model, using grid search over 3–10 maximum depths and $\{4k, 8k, 12k, 16k, 20k\}$ tree-ensemble sizes.

For propensity-related models, we trained feedforward neural networks with SiLU activations, width of 256 neuronal units, and depth of 3 layers. ADAM learning rates were 10^{-5} for $1,000,000$ epochs, with dropout of 5% and weight decay of 10^{-5}.

Policy learning. The `OT` program is employed with (1,000 nudges $\times$ 1,000 units) on the second/test sample per experimental invocation.

Dataset access. The Yelp dataset and its terms of use can be accessed at `https://www.yelp.com/dataset`.