Experimental Site Selection Under Directional Distribution Shifts

Adam Bouyamourn[*]

September 21, 2026

Abstract

Multi-site studies are widely used across the social sciences, education, and medicine to learn the effects of interventions across heterogeneous populations. But deployment populations, on which a given intervention is implemented, may differ systematically from the experimental population on which the intervention is evaluated. I propose a method for selecting sites with minimax guarantees against directional distribution shifts, which allows the researcher to guard against specified threats to the external validity of a multi-site study. I use Wasserstein Distributionally Robust Optimization to select sites with guarantees against directional shifts, provide a novel cutting-plane algorithm to implement the approach, and prove its optimality with respect to minimizing worst-case shifts. I highlight the empirical performance of the method with with two semi-synthetic simulations: I re-analyse the Tennessee STAR experiment, under the assumption that the study population underrepresented low-SES schools, and Crépon et al. (2015), which assesses the impact of increasing access to microcredit in rural Morocco, under the assumption that the study population was skewed towards villages with larger market potential. In these examples, I show that the WDRO method minimizes mean bias, average mean squared error, and W_1 distance under specified directional distribution shifts.

1 Introduction

Multi-site studies run a series of experiments across several locations or jurisdictions, with the goal of estimating treatment effects across experimental sites. They are widely used in the social sciences (Dunning et al., 2019; Blair et al., 2021; Banerjee et al., 2015), medicine (Rossouw et al., 2002; WHO Solidarity Trial Consortium et al., 2021), and education (Raudenbush and Schwartz, 2020; Orr et al., 2019).

They are a natural way to evaluate public policy interventions, which are often implemented at the level of geographic or institutional clusters, such as schools, hospitals, and legal jurisdictions. The effects of those interventions vary with the characteristics of the clusters,

[*]Assistant Professor and Faculty Fellow, Center for Data Science, New York University.

and with the composition of the subjects in them (Bloom et al., 2017; Raudenbush and Schwartz, 2020). Multi-site studies are popular because they have a suggestive ability to provide evidence about the *scope* of an empirical finding: whether an intervention produces results across many jurisdictions, whether it replicates, and whether its effects generalize or depend on context (Dunning et al., 2019; Bloom et al., 2017).

In practice, conclusions drawn from a given experimental sample may not carry over to the population of interest. When deployment populations on which a policy is enacted differ from the population on which an intervention was evaluated, estimates derived from the experimental sample may not accurately represent treatment effects in the target population, undermining the external validity and practical utility of research findings (Shadish et al., 2002; Findley et al., 2021; Egami and Hartman, 2023; Pearl and Bareinboim, 2011; Pearl and Bareinboim, 2014; Rudolph and Díaz, 2021; Egami and Hartman, 2021; Rudolph et al., 2024). This can be described as a problem of *distribution shift* from an observed experimental population to a deployment population of interest.

Distribution shift can take on a number of concrete forms. Characteristics and the composition of populations may change over time (Bansak et al., 2023; Rothenhausler and Bühlmann, 2023; Saville et al., 2022), subgroups of particular interest may be underrepresented (Tan et al., 2022; Hu et al., 2024), criteria for study inclusion may systematically misrepresent the population (Olsen and Orr, 2016; Olsen et al., 2022; Allcott, 2015), or data used to plan studies may be subject to measurement error (Bound et al., 2001). Composition of available experimental sites may differ systematically from the policy-relevant population on which the researcher wishes to experiment (Allcott, 2015), or minority groups may be systematically underrepresented in selected experimental units (Tan et al., 2022; Hu et al., 2024).

The forms of distribution shift to which a given study may be subject are often well-known to researchers in an applied area. In political science and economics, differences in institutional quality (Bold et al., 2018), trust in institutions (Cheeseman and Peiffer, 2023), rural-urban mix (Dehejia et al., 2019), and racial and ethnic context (Anoll et al., 2024; Hassell, 2021), can each be the source of substantive differences in the transportability of conclusions from one context to another.

In education, school districts enrolled in an experiment may be unrepresentative of the distribution of schools, leading to overly optimistic assessments of policy impacts (Olsen and Orr, 2016; Olsen et al., 2022). A specific controversy centers on the evaluation of the benefits of charter schools, in which the marginal effect of charter school attendance is compared between students who won a lottery to attend an oversubscribed charter school, and students who did not win the lottery, and instead attended a public school (Cohodes and Roy, 2024). These studies can conceal variation within the population of charter schools, however, by censoring the distribution of charter schools (Angrist et al., 2013).

In medicine, elderly patients and patients with common medical conditions are frequently excluded from randomized control trials (Van Spall et al., 2007; Zulman et al., 2011), while cancer trials systematicaly underenroll minority groups, among whom specific varieties of cancer are likely to be more prevalent (Oh et al., 2015). Women are also often frequently underenrolled in clinical trials (Daitch et al., 2022). These exclusions can lead to overly optimistic assessments of drug efficacy (Eichler et al., 2011), and of side effects in minority populations (Ramamoorthy et al., 2015).

This is analogous to the problem of selection bias in an observational study, where units opt into treatment on the basis of their expected returns from treatment (Heckman, 1974; Munro et al., 2025). A subtlety, however, is that treatment effect estimates are unbiased in the resulting multi-site trial, because a randomized experiment is conducted after the sites have been selected. The bias instead is in the gap between the experimental population and the population of interest, and cannot be diagnosed from the results of the experiment alone.

I propose a method for selecting sites in a multi-site study with guarantees against worst-case directional shifts in the deployment population. This allows the user to specify ex ante a threat to the external validity of a study by selecting a specific direction that maps onto a substantive threat to a study's transportability.

I first prove a new upper bound on the mean squared error of the population average treatment effect in the site selection problem, using optimal transport (Villani, 2008; Peyré and Cuturi, 2019). The bound decomposes the bias of a selection into two parts: bias due to observed site-level moderators, and bias due to unobserved moderators under the optimal transport plan on the observed ones. This gives a feasible target for optimization: minimize the transport distance between the site-level moderators of the selection and those of a hypothetical deployment population.

I write the planner's problem as a minimax problem. The planner chooses experimental sites to minimize the error of the selection against a set of adversarial distributions, generated by a model of distribution shift. I provide two parametric models of distribution shift: *secular drift*, which is modelled as translation with respect to a direction θ, and *selection bias*, which is modelled as an exponential tilt with weights that depend on θ.

I then use Wasserstein Distributionally Robust Optimization to choose a site selection with worst-case guarantees against the form of shift in the specified direction. I develop a novel cutting-plane algorithm that solves a minimax version of the site selection problem, and prove that it terminates with optimality guarantees on the resulting selection.

I reanalyse the Tennessee STAR experiment, and Crépon et al. (2015), a rural micocredit study in Morocco, under the assumption that their designs faced specific threats to external validity: in the former, that schools included in the study population were higher SES than schools in which the intervention would have been deployed; in the latter, that villages included in the study population had greater economic potential than villages that the policy would be deployed on. I show that the Wasserstein DRO method performs favorably in terms of mean bias and worst-case MSE under distribution shift compared to existing methods, which do not explicitly model the existence of adversarial distribution shifts.

I conclude with guidance on how to specify a direction and radius of robustness in practice, the need to select site-level covariates, and a suggestion to focus more resources on gathering information at the experimental planning stage about site-level features, and the composition of policy-relevant groups.

1.1 Related Literature

Recent work in site selection proposes methods to choose experimental locations that are designed with external validity in mind (Gechter et al., 2024; Egami and Lee, 2024; Olea et al., 2024).

Egami and Lee (2024) introduced explicit optimization methods for site selection in political methodology, and contributed to defining the problem of site selection. Their approach, based on the synthetic control method, uses optimization to select included sites that closely approximate sites that are not included in the selection, by estimating balancing weights (Abadie and Gardeazabal, 2003; Alberto Abadie and Hainmueller, 2010; Abadie and Zhao, 2025). Their goal is to have a high-quality weighted average representation of non-selected sites. In practice, this can be thought of as ensuring that non-selected sites are within the convex hull of selected sites.

Their formal result is an upper bound on the worst-case mean squared error of the predicted site-level effects under a smoothness condition on the effect surface. Given a target population of N sites with covariates, synthetic purposive sampling selects K sites and weights so that each non-selected site is approximated by a weighted average of the selected sites, in the manner of the synthetic control method (Abadie and Gardeazabal, 2003; Alberto Abadie and Hainmueller, 2010; Abadie and Zhao, 2025), with a penalty on the use of outlying sites. Its guarantee is an upper bound on mean squared error of the predicted site-level effects under a smoothness condition on the effect surface.

A key difference is that their solution set intends to solve for a convex hull that contains the PATE. This can lead to solution sets that include outliers. Their default setting is a regularize convex hull, which shrinks the convex hull by penalizing inclusion of outlying sites. This approach is motivated by practical concerns, but does not offer a formal guarantee on robustness to hypothetical deployment contexts.

Tipton (2013a) develops a subclassification estimator that reweights an existing experimental sample toward a population on strata of an estimated selection score; Tipton (2013b) turn the same idea into a design, stratifying the population of sites on the score or on clusters formed from the covariates and selecting sites within strata so that the sample's composition matches the population's across strata. The guarantee is that of stratified sampling: the sample effect is unbiased for the population effect when the effect surface is constant within strata, and its bias is bounded by within-stratum variation otherwise. The population is fixed and observed.

Olea et al. (2024) is the closest method to the current paper. They formulate site selection as a minimax decision problem: a policymaker chooses, for each site in a known set of policy sites, whether to adopt a treatment, and the effect surface is Lipschitz with a known constant. They show that selecting sites by solving the k-median problem solves a minimax-regret problem.

This paper draws on the **Distributionally Robust Optimization** (DRO) literature, a set of methods developed in operations research that find solution sets with guarantees against worst-case performance within the radius of a given solution (Ben-Tal et al., 2013;

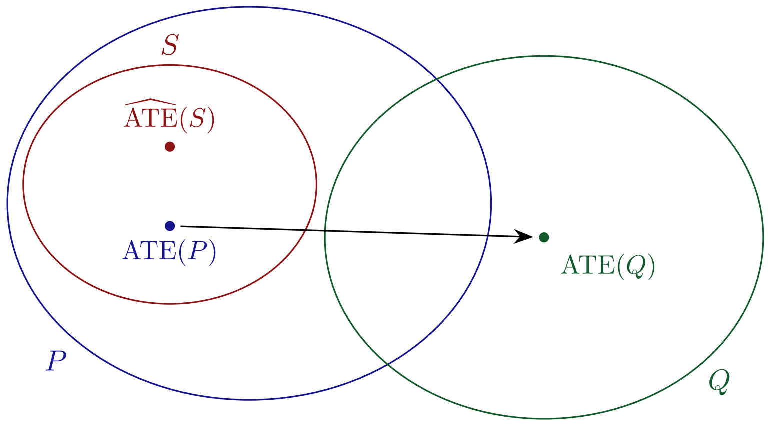
Esfahani and Kuhn, 2017; Kuhn et al., 2024; Blanchet et al., 2021; Blanchet and Murthy, 2019; Blanchet et al., 2024; Duchi and Namkoong, 2020; Levy et al., 2020; Bertsimas et al., 2023). DRO methods approach the problem of distributional uncertainty by providing statistical guarantees that a given solution is robust to a worst-case shift of the data. These methods provide insurance against poor performance within a specified neighborhood of the empirical solution (Luo and Mehrotra, 2020; Duchi and Namkoong, 2020). In statistics and causal inference, distributionally robust methods have been used for estimation under covariate shift and sample-selection bias: Duchi and Namkoong (2020) for uniform performance over subpopulations, Bertsimas et al. (2023) for causal effects from observational data, and Zhang et al. (2024) for generalizing heterogeneous effects across sites.

Minimax criteria for statistical decisions originate with Wald (1950). In treatment choice the relevant criterion is minimax regret (Manski, 2004; Stoye, 2009; Hirano and Porter, 2009). In experimental design, Kallus (2018) shows that the minimax-variance treatment assignment over a class of outcome functions is a balance criterion determined by that class, and Kallus (2021) characterizes when randomization is minimax; Bertsimas et al. (2015) optimize assignment directly; Hu et al. (2024) choose how many units to draw from each subpopulation to minimize worst-case regret over effect surfaces. This paper is closely related to the work of Kallus (2020) and Bertsimas et al. (2015), but considers a problem in which the data-generating process is assumed to be subject to distribution shift. Randomization is not necessarily minimax for this problem, and optimization of the nominal problem is not optimal for the shifted problem.

Section 2 develops the site selection problem under distribution shift, states bounds on the quantity of interest, and describes two parametric models of distribution shift. Section 3 motivates a nominal optimization problem, the cutting-plane algorithm used to solve the distributionally robust optimization problem, and states theoretical guarantees on the algorithm. Section 4 provides an assessment of extant site selection methods using data from the Tennessee STAR and (Crépon et al., 2015) microcredit experiments. Section 5 concludes.

2 Theory and Methods

2.1 The Site Selection Problem Under Distribution Shift



Let P be a set of candidate sites observed during the planning stage. The experimental planner's goal is to select a subset of sites $S \subset P$, where the subset is of size K, in order to minimize a realized loss on the deployed population of interest.

Each site s has associated covariates (X, U), where $X \in \mathbb{R}^d$ are observed at the planning stage, and $U \in \mathbb{R}^k$ are not observed. X is comprised of site-level covariates X_s, and aggregates of individual-level covariates $\bar{X}_i$, so that $X = \{X_s \cup \bar{X}_i\}$. We refer to these collectively as *site characteristics*, or site-level covariates, and do not distinguish between these elements in the sequel. Let P_X be the empirical distribution of observed site characteristics, P_U the distribution of unobserved site covariates, and $P_{X,U}$ their joint distribution. Likewise S_X, S_U and $S_{X,U}$ refer to their respective sample distributions.

We suppose that the policy being evaluated is enacted on a deployment population Q. We take Q to belong to a set of distributions $\mathcal{Q}_\rho$ generated from P_X by specific transformations corresponding to particular models of distribution shift (see Section 2.4). We suppose that these distributions are within a radius ρ of the original distribution P_X.

First, we define the Average Treatment Effect, as a function of an arbitrary distribution F which we take as the primary estimand of interest to the planner in a multi-site study.

Definition 2.1 (Average Treatment Effect)**.** Let $\tau : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ be the treatment effect as a function of covariates. Then, for a generic joint distribution F such that $(X, U) \sim F$,

$$ATE(F) = \int \tau(x, u)\, dF(x, u).$$

When F = P, this is the Population Average Treatment Effect, and when F = Q, this is the Average Treatment Effect in the Deployment Population. To avoid the ambiguity involved in 'PATE' under distribution shift, I prefer the more specific $ATE(P)$ or $ATE(Q)$ below. We define the **deployment error**, with respect to a deployment population Q:

Definition 2.2 (Deployment error)**.** The deployment error of a selection S with respect to a deployment population Q is:

$$\mathrm{MSE}_Q(S) = \mathbb{E}_{Z\sim\Psi, W\sim\Phi}\big[(\mathrm{ATE}(Q) - \widehat{\mathrm{ATE}}(S))^2\big].$$

Where the expectation is taken over the randomness in unit-level treatment assignment Z according to a design Ψ, and unit-level experimental inclusion W according to a sampling process Φ, and $\widehat{\mathrm{ATE}}(S)$ is the realized estimate of the ATE generated from running an experiment on the subset S.

In this paper, we are interested in minimizing the **worst-case deployment error** with respect to a set of distributions that are within a radius ρ of our observed distribution:

Definition 2.3 (Worst-case deployment error)**.** The worst-case deployment error of a selection S with respect to a deployment population Q within radius ρ of P_X is:

$$\sup_{Q\in\mathcal{Q}_\rho} \mathrm{MSE}_Q(S) = \sup_{Q\in\mathcal{Q}_\rho} \mathbb{E}\big[(\mathrm{ATE}(Q) - \widehat{\mathrm{ATE}}(S))^2\big].$$

The Researcher's goal is to minimize the worst-case deployment by choice of a subset S:

Problem 2.1 (Site selection under Distribution Shift)**.** Given candidate sites P, a budget K, and a set of deployment populations $\mathcal{Q}_\rho$, the Researcher's problem is:

$$\min_{S\subset P,\, |S|=K} \sup_{Q\in\mathcal{Q}_\rho} \mathrm{MSE}_Q(S).$$

The inner supremum is the deployment population that maximizes the deployment error of that choice, while the outer minimization describes the Researcher's choice of experimental sites.[1]

The Site Selection Problem Under Distribution Shift

1. **Planning.** The researcher defines the population of sites P and observes site-level covariates X for each candidate site. Given a budget constraint K, the researcher specifies a set of distribution shifts $\mathcal{Q}_\rho$ to guard against, and chooses a subset S.
2. **Experimentation.** An experiment is run in each selected site. Individual-level covariates are observed in the selected sites. Treatment is assigned within sites.
3. **Analysis.** The analyst estimates $\widehat{\mathrm{ATE}}(S)$, which also incorporates individual-level covariates, and may include with the declared design or welfare weights.
4. **Real-world Deployment.** A covariate distribution Q, and the deployment error, $MSE_Q(S)$, are realized.

Figure 1: The Researcher's Site Selection Problem

[1] In the setting without distribution shift, $\rho = 0$, $\mathcal{Q} = \{P_X\}$ and Problem 2.1 is $\min_S \mathrm{MSE}_{P_X}(S)$.

2.2 Assumptions

Problem 2.1 is infeasible without further assumptions on the relationship between the treatment effect function and covariates. We require that the treatment effect function is Lipschitz continuous in covariates (Lipschitz Continuity of τ), and that there is one treatment effect function common across units.

Assumption 2.1 (Lipschitz Continuity of τ)**.** *The treatment effect function $\tau : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ is Lipschitz continuous with constant L:*

$$|\tau(x, u) - \tau(x', u')| \leq L(\|x - x'\| + \|u - u'\|) \quad \forall x, x' \in \mathbb{R}^d, \forall u, u' \in \mathbb{R}^k$$

This ensures that treatment effects vary smoothly with covariates. When covariate values change, treatment effects must vary within an envelope defined by the size of the change of covariate values. This assumption is important, because it allows us to move from claims about covariates to claims about treatment effects.

We also require that the randomization design Ψ is independent of the optimal choice of subset:

Assumption 2.2 (Independence of Experimental Design and Site Selection)**.** *The randomization and unit sampling distributions in any given site Ψ_s, Φ_s do not depend on the choice of experimental subset.*

$$Z_s | S = A \sim \Psi_s \text{ and } W_s | S = A \sim \Phi_s, \ \forall A \subseteq P \text{ such that } s \in A.$$

Two useful properties follow from Assumption 2.2.

Lemma 2.1 (Unbiasedness of site-level estimation)**.**

$$\mathbb{E}_{Z\sim\Psi, W\sim\Phi}[\widehat{\text{ATE}}(S)] = \int \tau(x, u)\, dS_{X,U}(x, u) = \text{ATE}(S)$$

Define the site-specific estimation error $\text{Var}(\widehat{\text{ATE}}(S)) = \sigma_S^2$. Then:

Lemma 2.2 (Bias–variance decomposition)**.**

$$\text{MSE}_Q(S) = \big(\text{ATE}(Q) - \text{ATE}(S)\big)^2 + \sigma_S^2$$

The first term is controlled by choice of subset S, while the second term is controlled by choice of randomization design.

2.3 Bounds on the deployment error of a site selection

The assumptions above allow us to move from statements about covariates to statements about the error on the deployment distribution.

2.3.1 Technical Preliminaries

First, it is helpful to define the notion of a *transport plan*, which is a matrix that specifies a routing of mass from one distribution to another.

Definition 2.4 (Transport plan)**.** A **transport plan** between discrete distributions $P_X = \sum_{i=1}^{n} p_i \delta_{x_i}$ and $Q_Y = \sum_{j=1}^{m} r_j \delta_{y_j}$ is a matrix $\{\pi_{ij}\}_{(i=1,j=1)}^{(n,m)}$ with $\pi_{ij} \geq 0$ such that $\sum_{j=1}^{m} \pi_{ij} = p_i$ for each i and $\sum_{i=1}^{n} \pi_{ij} = r_j$ for each j.

Where δ is the Dirac delta. Row i of π sums p_i, the mass at P_{x_i}, while column j of π sums to r_j, the mass at Q_{x_j}.

In order to evaluate different transport plans, we need a way to assess the costs of a given proposed transport plan. A **cost function** describes the cost of travelling from X to Y. We use the Euclidean distance as our cost function, so that $c(X,Y) = \|X - Y\|_2 = \left(\sum_{l=1}^{d}(X_l - Y_l)^2\right)^{1/2}$.

The **optimal transport plan** is the plan π^* that in fact minimizes the distance between P and Q, for a given cost function $c(X,Y)$. That is,

Definition 2.5 (Optimal Transport Plan)**.** A transport plan π^* is optimal if

$$\pi^* = \arg\inf_{\pi} \sum_{i=1}^{n} \sum_{j=1}^{m} \pi_{ij} ||x_i - y_j||$$

That is, if π^* minimizes the cost of transporting mass from P to Q.

We can think of the solution to the optimal transport as being the shortest possible distance between X and Y, given the distributions P and Q.

Next, we define the 1-Wasserstein distance:

Definition 2.6. For two distributions μ and ν on $\mathbb{R}^d$ with finite first moments, the 1-Wasserstein distance is

$$W_1(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int \|x - x'\| \, d\gamma(x, x'),$$

where $\Gamma(\mu, \nu)$ is the set of couplings of μ and ν. For empirical distributions P and Q, the coupling γ is a *transport plan* (Kantorovich, 2006). γ^*, the transport plan that distributes mass from P to Q at infimal cost, is called the *optimal transport plan* (Villani, 2008).

Notice that the definition of W_1 distance is stated in terms of the optimal transport plan. Indeed, the W_1 distance from $P \to Q$ formalizes the notion of the the shortest distance from $P \to Q$ in the $\ell - 1$ norm: there are many ways to transport mass from P to Q, and W_1 is the distance associated with the lowest-cost way to do so.

2.3.2 Bounds on the deployment error

It is first helpful to consider the case where Q is known. In this case, we can derive the following upper bound on the deployment error:

Theorem 2.3 (Bound on the deployment error when Q is known and covariates are observed)**.** *Suppose Assumption 2.1 and Assumption 2.2 hold. Then, when all covariates are observed, for any selection S and a known deployment distribution Q,*

$$\mathrm{MSE}_Q(S) \leq L^2\, W_1(Q, S_X)^2 + \sigma_S^2,$$

The bound is infeasible as a design criterion, because Q is not observed when S is chosen. Indeed, our goal is to guard against the worst-case Q from a family of possible, unobserved shifts applied to our observed P_X.

To state a feasible bound, we want to decompose the distance between S and Q into observable and unobservable components. Recall that an optimal transport plan is a matrix that specifies the lowest-cost routing of mass from one distribution onto another. Let γ^* denote such a transport plan from S_X to Q_X. We consider the difference between the unobservable *conditional distributions* $S_{U|X}$, $Q_{U|X}$, under the optimal transport plan that maps S_X onto Q_X:

Definition 2.7 (Unobserved heterogeneity)**.**

$$\eta(S_U, Q_U) = \int W_1\big(S_{U|X=x},\, Q_{U|X=x'}\big)\, d\gamma^*(x, x').$$

This averages over the differences in unobservables across the selected set S and the deployment population Q. It is a measure of the sensitivity of a selection S to unobserved site-level heterogeneity in the deployment population Q.

We can then plug this quantity into our bound:

Theorem 2.4 (Bound on the deployment error with unobserved moderators)**.** *Let γ^* be an optimal coupling for (S_X, Q_X). Then, under Assumption 2.1 and Assumption 2.2*

$$\mathrm{MSE}_Q(S) \leq L^2\big[W_1(S_X, Q_X) + \eta(S_U, Q_U)\big]^2 + \sigma_S^2,$$

This bound decomposes the deployment error into an observable and unobservable component. We can therefore choose S to minimize the observable component of the bound, giving us the feasible problem:

Problem 2.2.

$$\min_{S \subset P,\, |S|=K}\; \sup_{Q \in \mathcal{Q}_\rho} W_1(S_X, Q_X).$$

$\mathcal{Q}_\rho$ is a set of distributions constructed from the observed distribution P_X. We next explain how these adversarial distributions are generated.

2.4 Directional Distribution Shift

We restrict our focus to specific varieties of distribution shift that are likely to arise in practice in applied policymaking contexts. In particular, we restrict our attention to *directional distribution shifts*.

Definition 2.8 (Directional distribution shift family)**.** Let $\mathscr{T} = \{T_\theta : \theta \in \Theta\}$ be a class of transformations from $\mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d)$, with $Q_\theta = T_\theta(P_X)$, a distribution induced by applying a specific transformation T_θ to the empirical distribution P_X. Define a restricted parameter space $\Theta_\rho = \{\theta \in \Theta : W_1(Q_\theta, P_X) \leq \rho\}$, that is, the set of values of θ for which the induced distribution is within a 1-Wasserstein radius of ρ of the original empirical distribution. Then, a distribution shift family of budget ρ indexed by parameter $\theta \in \Theta_\rho$ is a family of distributions:

$$\mathcal{Q}_\rho(\Theta) = \{T_\theta(P_X) : \theta \in \Theta_\rho, W_1(T_\theta(P_X), P_X) \leq \rho\}$$

Remark 1. *Each $\theta \in \Theta$ is a vector in $\mathbb{R}^m$, and its direction $\theta/\|\theta\|$ indexes a direction of shift. The set Θ therefore encodes which shifts the analyst regards as possible.*

When the set of possible deployment populations belongs to a directional shift family $\mathcal{Q}_\rho(\Theta)$, the problem becomes:

Problem 2.3 (DRO problem over a directional distribution shift family)**.**

$$\min_{S:|S|=K} \sup_{Q_\theta \in \mathcal{Q}_\rho(\Theta)} W_1(Q_\theta, S_X)$$

Where Q_θ now belongs to a directional distribution shift family, and the supremum picks out the worst-possible shift belonging to such a family. Choosing specific distribution shift families gives us robustness against particular models of distribution shift. These different versions of distribution shift have different substantive policy consequences, and we allow the analyst to decide which kind is relevant to their given their research goals and the application.

Model 2.1 (Secular drift (Translation))**.** Let $\Theta \subseteq \mathbb{R}^d$ be a set of possible directions along which the covariates may drift, and let T_θ shift every site by the vector θ:

$$Q_\theta = \frac{1}{|P|} \sum_{s=1}^{|P|} \delta_{X_s + \theta}.$$

For instance, deployment sites could be systematically poorer, more rural, or larger than the planning data, by up to ρ standard deviations.

Model 2.2 (Selection bias (Exponential tilting))**.** Let $\Theta \subseteq \mathbb{R}^d$ be a set of directions along which sites may be over- or under-represented, and let T_θ reweight the sites log-linearly:

$$Q_\theta = \sum_{s=1}^{|P|} w_s(\theta)\, \delta_{X_s}, \qquad w_s(\theta) = \frac{\exp\{\theta^\top X_s\}}{\sum_{r=1}^{|P|} \exp\{\theta^\top X_r\}}.$$

This model of distribution shift is applied in Tibshirani et al. (2020), and mirrors the problem of site selection bias described in Allcott (2015). Suppose that sites on the candidate list were non-randomly sampled from a deployment population, so that each site s had sampling weight $\psi_s \neq \frac{1}{|P|}$. In order to reweight the candidate population P_X back to the original deployment population Q, we would want to solve for Horvitz-Thompson weights $q_s \propto \frac{|P|^{-1}}{\psi_s}$. If we suppose that the weights $\psi_s \propto \exp\{-\theta^\top X_s\}$, we get the exponential tilting model above.

3 Algorithms and Computation

3.1 Warm-up: Optimal Transport and the Site Selection Problem

To motivate the distributionally-robust optimization problem used to solve the site selection problem, it is helpful first to consider the simpler optimization problem when the deployment population is the observed population P_X.

Problem 3.1 (Optimization problem over an observed distribution)**.**

$$\min_{S:|S|=K} W_1(P_X, S_X)$$

It is helpful to substitute into this problem the definition of the discrete 1-Wasserstein distance, to get the following problem statement:

Problem 3.2 (Optimization problem over an observed distribution)**.**

$$\min_{S\,:\,|S|=K} \min_{\gamma\in\Gamma(P_X,S_X)} \sum_{j=1}^{|P|}\sum_{k\in S}\gamma_{jk}\,\|X_j^P - X_k^S\|,$$

This problem has an inner minimization step, and an outer minimization step. These are solved simultaneously as a Mixed Integer Linear Program, but it is helpful to provide intuition for both steps.

The inner minimization step is an optimal transport problem. It takes a fixed selection of sites $\tilde{S}$, and finds an optimal transport plan $\gamma^*(\tilde{S})$, that minimizes the distance between P_X and $\tilde{S}_X$, given $\tilde{S}$ fixed. Because it is an optimal transport problem, it can be stated as a linear program (Kantorovich, 2006; Dantzig, 1951).

The outer minimization step chooses the set S^* that minimizes the total objective, given the optimal transport plan $\gamma^*(S)$ for each possible subset S. This step adds integer constraints, because site selection is a discrete selection problem. Selecting sites given fixed losses is a Mixed Integer Problem. The combined problem is therefore a Mixed Integer Linear Program.

The combined problem is solved by branch-and-bound (Land and Doig, 1960; Dakin, 1965). First, the integer constraints are relaxed, reducing the MILP to a Linear Program with fractional site assignments. Then, the solver evaluates the objective for each pair of subproblems, with the fractional assignments fixed to integer constraints. A subproblem is discarded when its relaxation is dominated by the current best integer solution.

The solution set S^*_{MILP} generated from this procedure minimizes the observed distance between S_X and P_X, which is the case when distribution shift is assumed not to occur, the deployment population is the observed population, and $\rho = 0$.

3.2 Cutting-plane algorithm for site selection under distribution shift

Now recall our optimization problem of interest:

For a target distribution P_X and selection distribution S_X:

$$\min_{s,\pi} \quad \sum_{j=1}^{|P|}\sum_{k=1}^{|P|} \pi_{jk} \, \|X_j^P - X_k^S\|$$

subject to:

$$\sum_{k=1}^{|P|} s_k = K \qquad \text{(Site budget)}$$

$$\sum_{k=1}^{|P|} \pi_{jk} = \frac{1}{|P|} \quad \forall j \qquad \text{(Target marginal)}$$

$$\sum_{j=1}^{|P|} \pi_{jk} = \frac{s_k}{K} \quad \forall k \in P \qquad \text{(Selection marginal)}$$

$$\pi_{jk} \geq 0 \quad \forall j, k \in P \qquad \text{(Nonnegativity)}$$

$$s_k \in \{0, 1\} \quad \forall k \in P \qquad \text{(Binary selection)}$$

Figure 2: MILP formulation of the Site Selection Problem

Problem 3.3 (DRO problem over a directional distribution shift family)**.**

$$\min_{S:|S|=K} \sup_{Q_\theta \in \mathcal{Q}_\rho(\Theta)} W_1(Q_\theta, S_X)$$

It is instructive to note that this is a minimax problem, and we solve it by implementing a cutting-plane algorithm.

This problem requires us to generate a set of adversarial distribution shifts in observed covariates according to the model of distribution shift chosen by the analyst:

$$Q_{\theta^{(1)}}, Q_{\theta^{(2)}}, \cdots \in \mathcal{Q}_\rho(\Theta)$$

We specify a mesh of width h over the parameter grid $\mathcal{G}_h \subset \Theta$. This generates a grid of parameter values $\mathcal{G}_h = \theta_1, \ldots, \theta_{|\mathcal{G}_h|}$, and we can generate one distribution Q for each point on this grid. We refer to this grid of shifted distributions as $\mathcal{Q}_h$.

The cutting-plane algorithm requires three subroutines. The first, EvaluateSiteSelectionUnderQ(S, Q), takes a proposed site selection $\tilde{S}$, and finds an optimal transport plan between $\tilde{S}$ and Q. This is a linear program, because the site selection is fixed. We use this as a general utility to evaluate the cost of different adversarial distributions Q with respect to a proposed site selection $\tilde{S}$. ($\tilde{\cdot}$ is used to denote an intermediate selection of the procedure.)

The next is the adversary's move, AdversaryBestResponse$(\tilde{S})$. This takes a proposed site selection $\tilde{S}$ and evaluates it against every distribution shift generated along the grid $\mathcal{Q}_h$. It yields a vector of costs $W_1(Q, \tilde{S})$, for each Q in $\mathcal{Q}_h$. Taking the maximum cost,

Algorithm 1 Cutting-plane method for site selection under distribution shift

Input: Candidate covariates $X_1^P, \ldots, X_{|P|}^P$; budget K; shift family $\mathcal{Q}_\rho(\Theta)$; mesh h; nominal selection S^*_{MILP}; tolerance $\epsilon \geq 0$

Output: Robust selection S^*_{DRO}

---

procedure EvaluateSiteSelectionUnderQ(S, Q)

  Returns: $W_1(S, Q)$, the cost of an optimal transport plan from $S \to Q$

  $C_{js} \leftarrow \|X_j^Q - X_s^P\|$ for $j \leq |P|$, $s \in S$

  $\tilde{\pi} \leftarrow \arg\min \left\{ \sum_j \sum_{s \in S} \pi_{js} C_{js} \ : \ \pi \geq 0, \quad \sum_{s \in S} \pi_{js} = w_j \ \forall j, \quad \sum_j \pi_{js} = \frac{1}{K} \ \forall s \in S \right\}$

  return $\sum_{j,s} \tilde{\pi}_{js} C_{js}$

procedure AdversaryBestResponse($\tilde{S}$)

  Returns: Highest cost adversarial shift $\tilde{Q}$ over a grid, given a site selection $\tilde{S}$

  for $Q \in \mathcal{Q}_h$ **do**

    $W_1(Q, \tilde{S}_X) \leftarrow$ EvaluateSiteSelectionUnderQ($\tilde{S}, Q$)

  $\tilde{Q} \leftarrow \arg\max_{Q \in \mathcal{Q}_h} W_1(Q, \tilde{S}_X)$

  return $\tilde{Q}$

procedure PlannerBestResponse($\tilde{\mathcal{Q}}$)

  Returns: A selection $\tilde{S}$ that minimizes the error induced by the worst shift $\tilde{Q}$ in a collection $\tilde{\mathcal{Q}}$ via MILP

  $C_{js}^{(Q)} \leftarrow \|X_j^Q - X_s^P\|$ for $Q \in \tilde{\mathcal{Q}}$, $j \leq |P|$, $s \leq |P|$

$$\tilde{S} \leftarrow \arg\min_{\mathbb{I}\{s \in \tilde{S}\} \in \{0,1\},\ \gamma^{(Q)} \geq 0,\ t \in \mathbb{R}} \quad t$$

$$\begin{aligned} \text{subject to} \quad & t \ \geq \ \textstyle\sum_{j,s} \gamma_{js}^{(Q)} C_{js}^{(Q)} && \forall Q \in \tilde{\mathcal{Q}}, \\ & \textstyle\sum_s \mathbb{I}\{s \in \tilde{S}\} = K, \\ & \textstyle\sum_s \gamma_{js}^{(Q)} = w_j^{(Q)} && \forall j, \ \forall Q \in \tilde{\mathcal{Q}}, \\ & \textstyle\sum_j \gamma_{js}^{(Q)} = \mathbb{I}\{s \in \tilde{S}\}/K && \forall s, \ \forall Q \in \tilde{\mathcal{Q}} \end{aligned}$$

  return $\tilde{S}$

$\mathcal{Q}_h \leftarrow \{Q_\theta : \theta \in \mathcal{G}_h\}$, where $\mathcal{G}_h \subset \Theta_\rho$ is a grid of mesh h with $0 \in \mathcal{G}_h$

$\tilde{\mathcal{Q}}^{(0)} \leftarrow \{P_X\}$; $\tilde{S}^{(0)} \leftarrow S^*_{\text{MILP}}$; $t \leftarrow 0$

repeat

  $\tilde{Q}^{(t+1)} \leftarrow$ AdversaryBestResponse($\tilde{S}^{(t)}$)

  $\tilde{\mathcal{Q}}^{(t+1)} \leftarrow \tilde{\mathcal{Q}}^{(t)} \cup \{\tilde{Q}^{(t+1)}\}$ ▷ Worst-case shift added to collection

  $\tilde{S}^{(t+1)} \leftarrow$ PlannerBestResponse($\tilde{\mathcal{Q}}^{(t+1)}$)

  $\text{UB}^{(t)} \leftarrow$ EvaluateSiteSelectionUnderQ($\tilde{S}^{(t)}, \tilde{Q}^{(t+1)}$)

  $\text{LB}^{(t+1)} \leftarrow \max_{Q \in \tilde{\mathcal{Q}}^{(t+1)}}$ EvaluateSiteSelectionUnderQ($\tilde{S}^{(t+1)}, Q$)

  $t \leftarrow t + 1$

until $\min_{u<t} \text{UB}^{(u)} - \text{LB}^{(t)} \leq \epsilon$

return $S^*_{\text{DRO}} \leftarrow \tilde{S}^{(u^\star)}$ with $u^\star \in \arg\min_{u<t} \text{UB}^{(u)}$

For any observed sites selection $\tilde{S}$, the adversary can evaluate the loss of $\tilde{S}$ over the whole grid of generated shifts, and return whichever shift $\tilde{Q}$ generated the largest observed optimal transport plan cost.

It also requires us to solve a version of the MILP above for a set of observed adversarial distributions. This helper function is called $\textsc{PlannerBestResponse}(\tilde{Q})$. In particular, we want to solve the MILP for the worst-case adversarial distribution. To do this, we specify the epigraph of the loss due to the site selection, t, and set it greater than or equal to the largest cost of the worst $\tilde{Q}$ observed in a specific set $\tilde{\mathcal{Q}}$. This formalization trick allows us to minimize a linear program against the maximum loss, which is nonlinear.

The algorithm then proceeds as follows. First, the algorithm is initialized using the solution set S^*_{MILP} from Section 3.1, and a grid of parameter values is generated. Second, the adversary evaluates the cost of a set of optimal transport plans from S^*_{MILP} to the set of adversarial distributions $\mathcal{Q}_h$, and selects the highest-cost adversarial distribution $\tilde{Q}$. This worst-case distribution is added to a storage set of adversarial distributions $\tilde{\mathcal{Q}}$. Third, the planner chooses a best response to the worst-case distribution in the set $\tilde{\mathcal{Q}}$ via MILP.

The value of the worst-cost transport plan from the previous selection, evaluated before the planner chooses a new set of sites, here, S^*_{MILP}, is stored as the upper bound (UB) on the loss: it is the worst-case loss of a feasible selection. The value of the worst-cost transport plan from the current selection is stored as the lower bound in the loss: the adversary can only increase the loss by re-selecting an adversarial distribution.

We therefore have a nonincreasing minimum upper bound on the worst-case transport loss, and a nondecreasing lower bound on the worst-case transport loss. The algorithm terminates when $\min\limits_{u<t} UB^{(u)} - LB^{(t)} \leq \epsilon$, at which point the worst-case transport loss over the parameter grid is within epsilon of its optimal value, and the site selection is output as S^*_{DRO}.

3.3 Guarantees

Theorem 3.1 (Guarantee for Algorithm 1)**.** *Suppose Θ_ρ is bounded, $\mathcal{G}_h \subset \Theta_\rho$ is a grid such that every $\theta \in \Theta_\rho$ lies within h of some grid point, and* $\textsc{PlannerBestResponse}$ *returns an exact minimizer. Then Algorithm 1 terminates after at most $|\mathcal{Q}_h|$ iterations, and its output S^*_{DRO} satisfies*

$$\sup_{\theta \in \Theta_\rho} W_1\big(Q_\theta, S^*_{\text{DRO},X}\big) \;\leq\; \min_{S:|S|=K} \; \sup_{\theta \in \Theta_\rho} W_1\big(Q_\theta, S_X\big) \;+\; \epsilon \;+\; c\,h,$$

Where c is a Lipschitz constant.

4 Empirical Studies

4.1 Tennessee STAR: Undersampling of low-SES schools

Simulation Design

The Tennessee STAR experiment studied the effect of smaller class sizes on learning outcomes at 75 schools. It found that smaller class sizes improved test scores (Mosteller, 1995).

A plausible threat to validity would occur if schools enrolled in the STAR experiment were more socioeconomically advantaged than schools in which the intervention was eventually implemented.

Because we observe outcomes and covariates for 75 schools, we can do two things. First, we can partition schools into a *candidate set* and a *deployment set*. The candidate set is observed by the methods, and used to make a decision about where to experiment. We can manipulate the deployment set according to a model of distribution shift in order to induce distribution shift in the held-out sites.

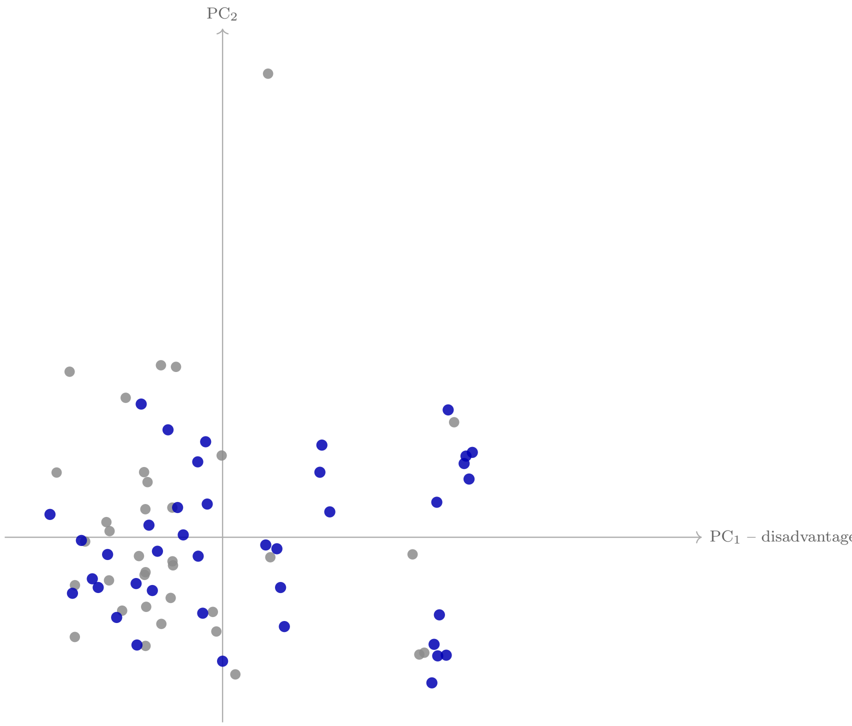
We first project covariates into principal component space. The first principal component loads on the covariates *free-lunch share*, *minority share*, and weakly on *enrollment*. We interpret this as a dimension of socioeconomic disadvantage. For the purpose of the method, we ask, what would happen to the estimate of the PATE if the policy were deployed in a population more disadvantaged than the average in the experimental population?

We can formalize this as a *secular shift* in the positive direction of the first principal component. The direction θ is chosen so that the shift occurs along PC1. I vary its extent through simulation, and the norm of θ describes the size of the shift in terms of standard deviations along PC1.

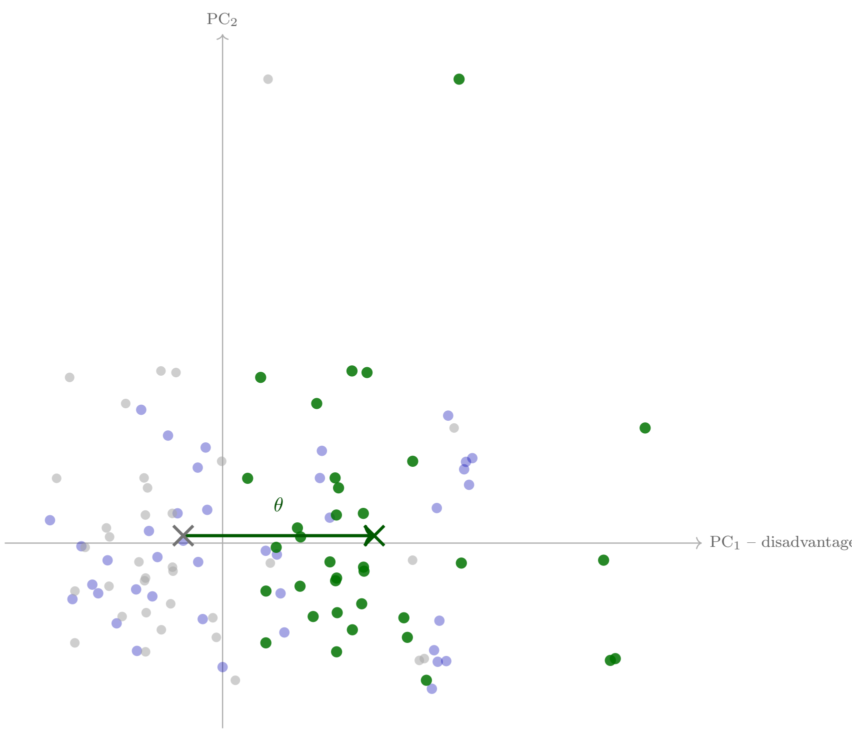
To summarize, our approach is as follows.

Table 1: Simulation design for the STAR re-analysis.

Step	Description
1. Candidate set	Draw 40 schools at random from the 75 STAR schools.
2. Selection	Each method selects K schools from the candidate set.
3. Deployment set	Shift the 35 held-out schools by θ^*, a translation of $\|\theta^*\|$ standard deviations along the first principal component. The shifted schools are the deployment population Q.
4. Estimate	Run the experiment in the selected schools, at their candidate-set covariates. Compare the estimate of the PATE from the selected sites to the PATE in Q.
5. Report	Mean absolute bias of the PATE estimate, its mean squared error, and the mean W_1 distance from the selected sites to the deployment population



(a) Candidate set (blue, $n = 40$) and held-out set (grey, $n = 35$).



(b) Translation shift applied to the held-out schools. Crosses mark the shift from $ATE(P)$ to $ATE(Q_\theta)$.

Figure 3: Simulation design. The sites are randomly partitioned into candidate sites and held-out sites (top). Held-out sites are then shifted adversarially in the direction of θ (bottom).

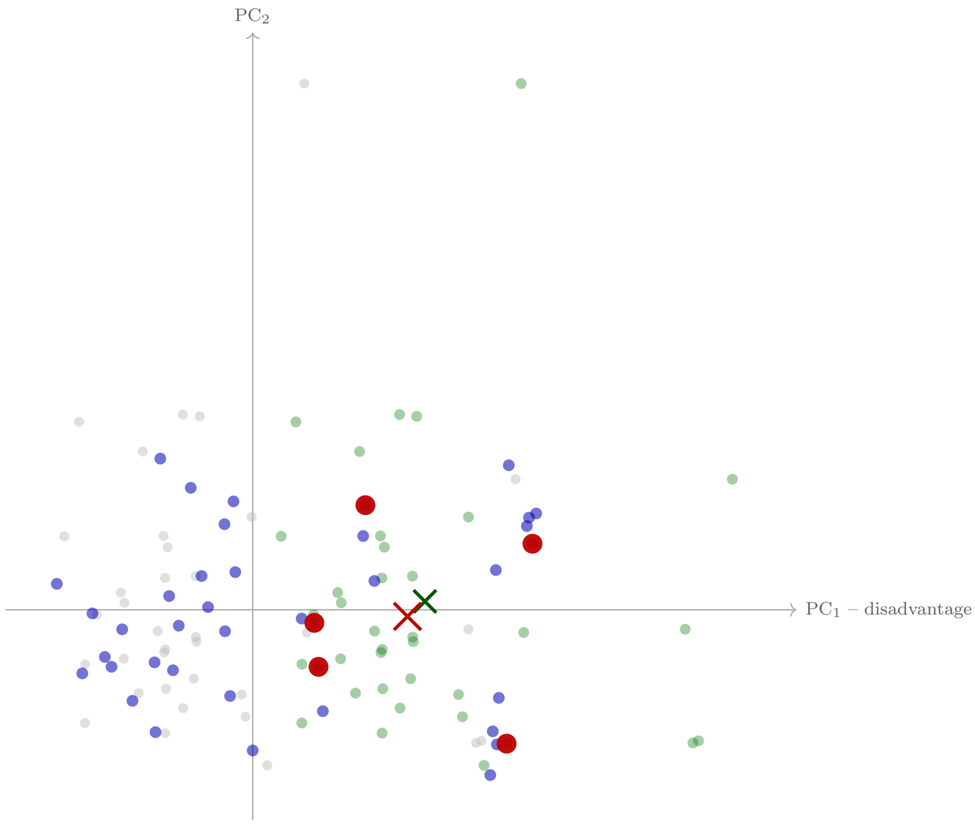


Figure 4: Sites selected by the Wasserstein DRO method under distribution shift. Green cross indicates $ATE(Q_\theta)$, red cross indicates $ATE(S^*_{\text{DRO}})$

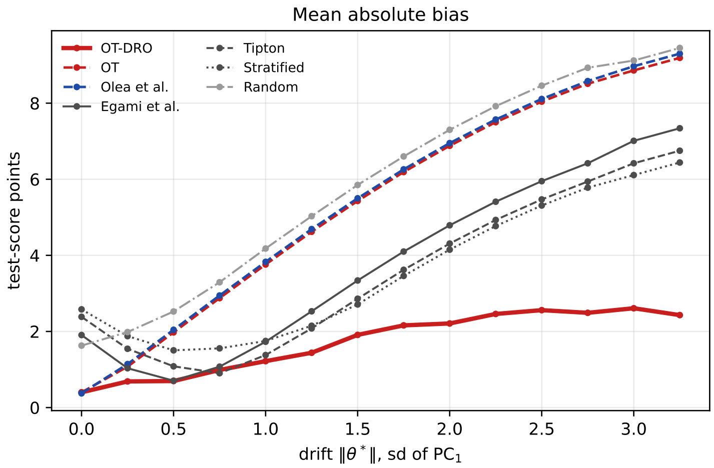
Results

In the above simulation, the Wasserstein-DRO method performs favourably under induced distribution shift, whenever distribution shift is greater than zero. It does so for the quantity it is explicitly intended to target, the W_1 distance, but also for the MSE and bias as well, which it upper bounds.

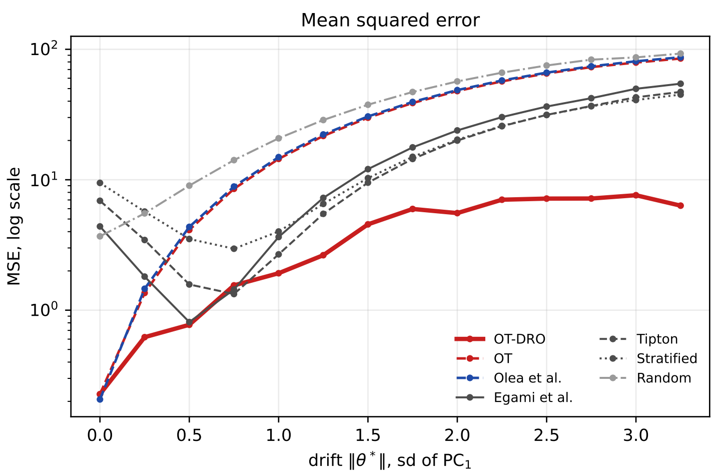
The method of Olea et al. (2024) outperforms WDRO in the absence of distribution shift. This is in part because WDRO pays 'the price of robustness' (Bertsimas and Sim, 2004) – a method optimal for the robust counterpart of an optimization problem is generally not optimal for the nominal version of that optimization problem.

The method has a structural advantage in this simulation, in that it 'knows' the correct value of θ.

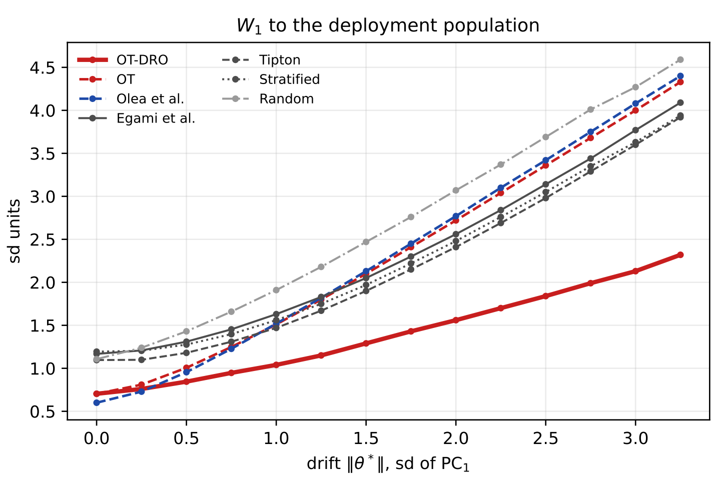
Further simulation studies are needed to assess the method i) when θ is misspecified ii) under different levels of covariate informativeness and ii) under different models of distribution shift.



(a) Mean absolute bias of the deployment estimate.



(b) Mean squared error of the deployment estimate, log scale.



(c) Mean W_1 distance from the selected sites to the deployment population.

Table 2: Site selection under covariate drift

(a) Mean absolute bias of the deployment estimate, test-score points.

method	$\|\theta^*\| = 0$	1	2	3
W-DRO	0.40	**1.22**	**2.21**	**2.61**
Olea et al.	**0.37**	3.83	6.95	8.97
Egami et al.	1.91	1.73	4.79	7.01
Tipton	2.38	1.38	4.31	6.42
Stratified	2.58	1.75	4.15	6.11
Random	1.63	4.18	7.30	9.12

(b) Mean squared error of the deployment estimate.

method	$\|\theta^*\| = 0$	1	2	3
W-DRO	**0.2**	**1.9**	**5.5**	**7.6**
Olea et al.	**0.2**	14.9	48.7	80.9
Egami et al.	4.4	3.6	23.9	49.8
Tipton	6.9	2.7	20.0	42.8
Stratified	9.5	4.0	20.3	40.9
Random	3.7	20.8	56.7	86.7

(c) Mean W_1 distance from the selected sites to the deployment population

method	$\|\theta^*\| = 0$	1	2	3
W-DRO	0.70	**1.04**	**1.56**	**2.13**
Olea et al.	**0.60**	1.52	2.77	4.08
Egami et al.	1.17	1.63	2.56	3.77
Tipton	1.10	1.47	2.41	3.60
Stratified	1.19	1.56	2.48	3.63
Random	1.11	1.91	3.07	4.27

4.2 Crépon et al.: overweighting of high market potential villages

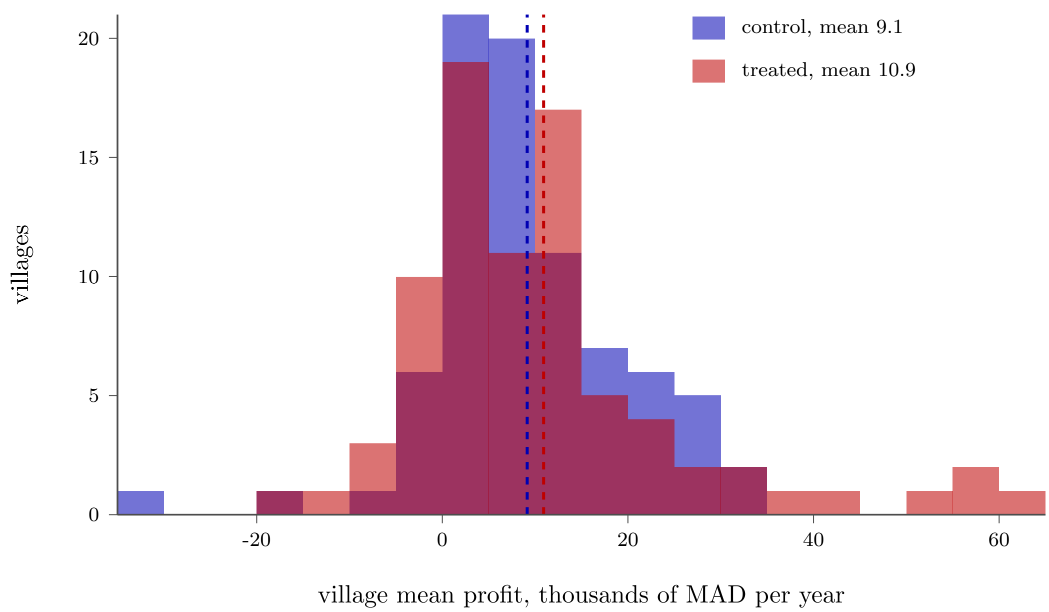
Study Characteristics

Crépon et al. (2015) studied the effect of an expansion in the availability of micocredit in rural areas in Morocco, when microfinance institution Al Amana expanded into a number of villages in 2006, on local farm profits and poverty reduction. The study population consisted of 162 villages in 81 matched pairs, where one village in each pair was randomly assigned to receive a new bank branch, which issued loans of a value between 1000 and 15000 Moroccan Dirhams (\$105 to \$1500). We take the matched pair as the experimental site. Each branch's service area was fixed by a map, shown in Figure 8. Loan take-up was 13 % in treated villages, and 0% in untreated villages.

The loans were found to increase self-employed profits by 2006 MAD against a control mean of 9056 MAD, an increase of around 22%. Profit effects were heterogeneous across villages.

In this exercise, we ask, what would the results of this study have been if the inclusion mechanism overweighted the market potential of the villages in the study? This is intended to recreate the problems described by Allcott (2015), in which a site selection yields an overly optimistic estimate of the effect of a policy.

Figure 6: Heterogeneous profits across villages.



Simulation Design

To build an adversarial situation for this study, we project site-level covariates onto principal components, and construct a reweighting shift that upweights villages with lower market potential. Lower market potential includes the covariates: baseline business, agricultural and livestock activity, business assets, asset index, prior borrowing, and poverty (negatively signed).

For this exercise we use the *tilt* model of drift, which is a reweighting shift in the negative direction of market potential. Figure 7 shows that the applied transformation tilts the distribution of site-level profits to the left.

The estimate of the ATE(Q) under $||\theta|| = 1$ is -723 and -2252 under $||\theta|| = 2$, indicating that these shifts reverse the sign of the paper's main finding.

Figure 7: Tilted profits.

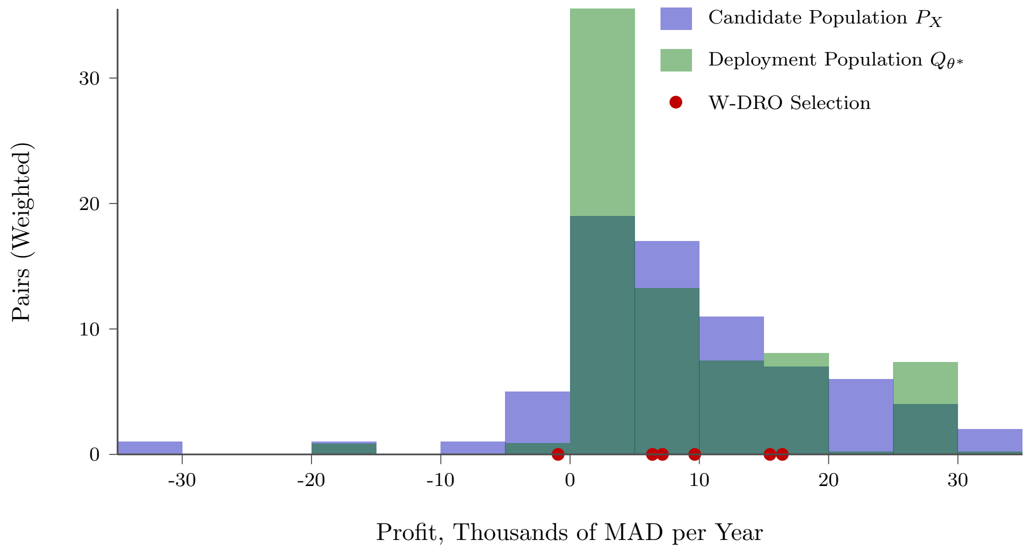
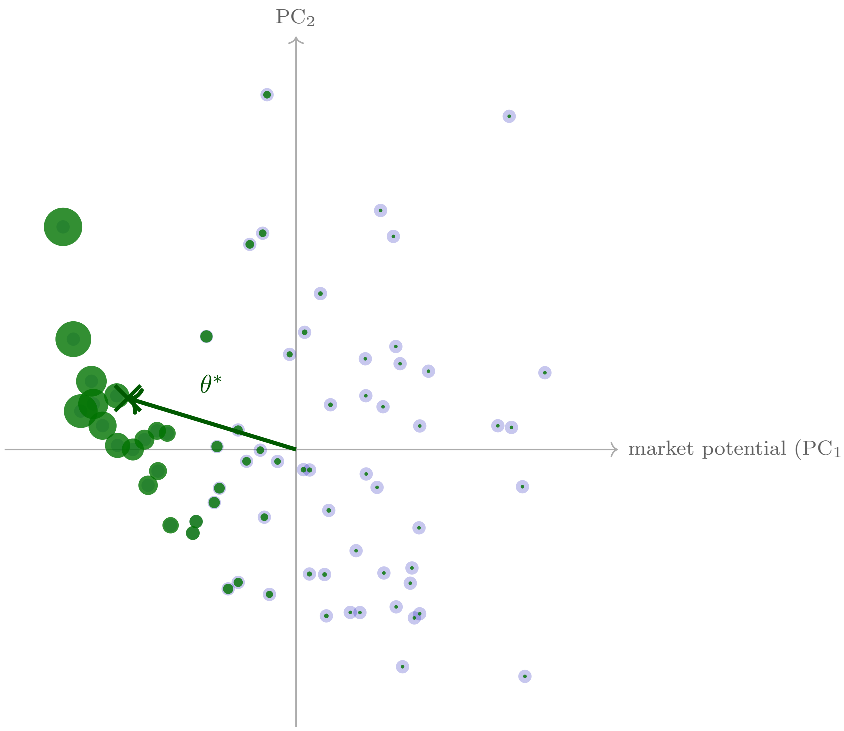


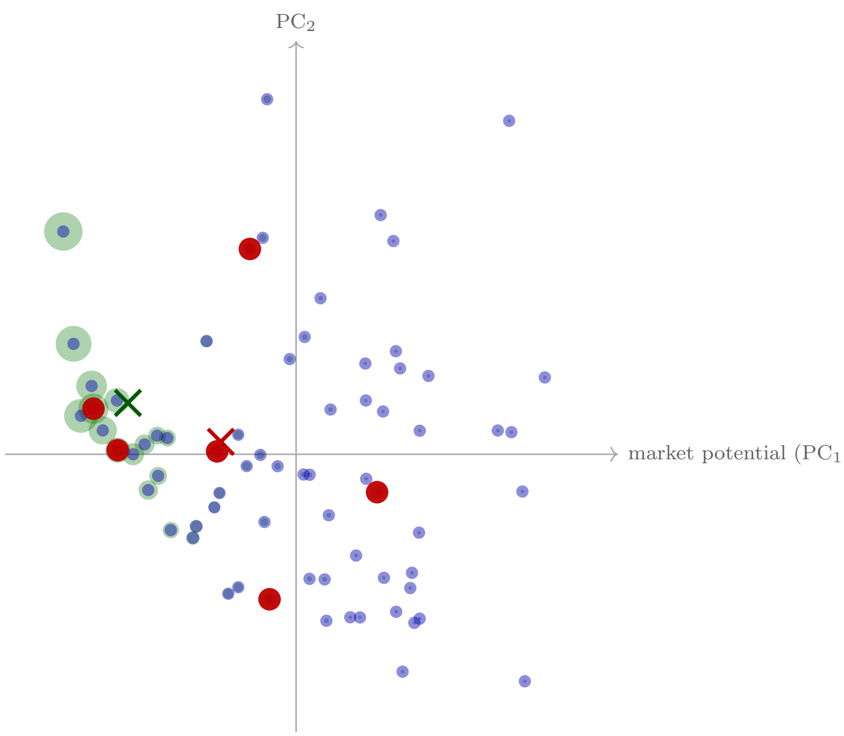
Table 3: Simulation design for the Crépon re-analysis.

Step	Description
1. Sites	The matched village pairs, with site-level effects estimated as the treated–control difference in mean profit within each pair.
2. Candidate set	All pairs, each with equal weight. Site selection methods observe the pair-level baseline covariates.
3. Selection	Each method selects $K = 6$ pairs. The Wasserstein DRO selection hedges over tilts toward low market potential, up to $\rho = \|\theta^*\|$.
4. Deployment set	The same pairs reweighted by $q_s \propto \exp(\theta^{*\prime} X_s)$, with θ^* pointing toward low market potential at $\|\theta^*\| \in \{0, 1, 2\}$ standard deviations of that index.
5. Estimate	Compare the equal-weight estimate from the selected pairs to the PATE under the deployment weights.
6. Report	Mean absolute bias, mean squared error, and mean W_1 distance from the selected sites to the deployment population. Deterministic methods appear once; stratified and random selection average 100 draws.

Figure 8: Branch intervention areas in the Crépon et al. study.



(a) Tilted sites in the direction of low market potential



(b) The Wasserstein DRO selection (red), with the ATE(S), ATE(Q) indicated.

Figure 9: The Crépon design under a tilt toward low market potential, $\|\theta^*\| = 2$.

Results

Table 4: Site selection under a tilt toward low market potential: self-employment profit.

(a) Mean absolute bias, dirhams per year.

method	$\|\theta^*\| = 0$	1	2
W-DRO	**82**	**2181**	**1660**
Olea et al.	84	2517	3611
Egami et al.	244	2677	3771
Tipton	2073	4506	5599
Stratified	1239	3445	4538
Random	1332	2643	3716

(b) Mean squared error, millions of squared dirhams.

method	$\|\theta^*\| = 0$	1	2
W-DRO	**0.01**	**4.76**	**2.76**
Olea et al.	0.01	6.34	13.04
Egami et al.	0.06	7.17	14.22
Tipton	4.30	20.30	31.35
Stratified	2.31	13.15	21.88
Random	2.53	9.37	16.30

(c) Mean W_1 distance to the deployment population, standard-deviation units.

method	$\|\theta^*\| = 0$	1	2
W-DRO	1.068	**1.120**	**1.219**
Olea et al.	**0.994**	1.322	1.714
Egami et al.	2.275	2.663	3.015
Tipton	1.252	1.676	2.066
Stratified	1.698	2.118	2.492
Random	1.494	1.639	1.939

4.3 Discussion

In each case, the WDRO method described above performs best under distribution shift under a specified direction. This is because it is design to protect against the particular adversarial case constructed in the simulation example. This is intended to validate the method: when deployment risks are known, the WDRO method effectively hedges against those risks, in terms of minimizing the expected error of downstream estimands as a function of unobserved drift.

Several qualifications are worth making. First, these simulations presuppose that the direction of the shift is known in advance. This is unrealistic in practice. But the purpose of using these methods is to generate designs that are conservative with respect to specified threats to external validity. Thus, using the method gives you a guarantee of conservatism against possible threats: that is the nature of the statistical guarantee, and is valuable for a study designer who has not yet run a study. Second, there is a clear role for substantive knowledge to guide the generation of potential threats. This creates researcher degrees of freedom, but the adjustment is nonetheless conservative, as it allows the researcher to pre-empt possible threats to external validity and guard against them.

A third point is that the other methods assessed here are not directional: they hedge in every possible direction simultaneously. This means that they conservatively protect against many possible directions of shift. In practice, researchers are likely to believe that some directions are worse than others in terms of the significance of results.

5 Conclusions

I develop a novel method for site selection that provides guarantees against directional distribution shifts during the planning stage. This method is useful for practitioners in applied contexts designing multi-site studies when specific threats to external validity are known. This method then allows researchers to choose sites with minimax guarantees against specific threats to external validity.

This method is intended to address the existence of well-documented threats to the validity of large-scale policy evaluations, where site selection bias threatens to lead to biased or inaccurate conclusions drawn from randomized control trials (Allcott, 2015; Olsen et al., 2022).

This method could be encouraged by funding bodies with specific objectives in terms of ensuring representation of minority groups in particular trials.

The utility of optimization-based methods in general requires that covariates are collected at the planning stage, and are informative. How can researchers assess the relationship between covariates and expected effects? They can refer to prior experiments; assess the relationship between covariates and pre-treatment outcomes; refer to estimated effects from pilot studies or studies in the same substantive domain; and use domain expertise and substantive knowledge to make suppositions grounded in a mechanistic understanding of the causal relationship.

Two rhetorical aims of this project is to encourage funding bodies that support multi-site studies to 1) explicitly state threats to external validity ex ante, and commit to experimental designs that pre-empt these threats 2) to commit to collecting informative covariates about trial populations in advance of conducting studies.

In future work, I intend to extend this project to optimal enrollment of individuals for estimation of the CATE under distribution shift, and to impose structure on θ to allow for unit enrollment decisions with guarantees on the Conditional Value-at-Risk of an intervention on the welfare of specified subgroups.

References

Abadie, A. and J. Gardeazabal (Mar. 2003). "The Economic Costs of Conflict: A Case Study of the Basque Country". In: *American Economic Review* 93.1, pp. 113–132. DOI: 10.1257/000282803321455188. URL: https://www.aeaweb.org/articles?id=10.1257/000282803321455188.

Abadie, A. and J. Zhao (2025). *Synthetic Controls for Experimental Design*. arXiv: 2108.02196 [stat.ME]. URL: https://arxiv.org/abs/2108.02196.

Alberto Abadie, A. D. and J. Hainmueller (2010). "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program". In: *Journal of the American Statistical Association* 105.490, pp. 493–505. DOI: 10.1198/jasa.2009.ap08746. eprint: https://doi.org/10.1198/jasa.2009.ap08746. URL: https://doi.org/10.1198/jasa.2009.ap08746.

Allcott, H. (Mar. 2015). "Site Selection Bias in Program Evaluation *". In: *The Quarterly Journal of Economics* 130.3, pp. 1117–1165. ISSN: 0033-5533. DOI: 10.1093/qje/qjv015. eprint: https://academic.oup.com/qje/article-pdf/130/3/1117/30637203/qjv015.pdf. URL: https://doi.org/10.1093/qje/qjv015.

Angrist, J. D., P. A. Pathak, and C. R. Walters (Oct. 2013). "Explaining Charter School Effectiveness". In: *American Economic Journal: Applied Economics* 5.4, pp. 1–27. ISSN: 1945-7790. DOI: 10.1257/app.5.4.1. URL: http://dx.doi.org/10.1257/app.5.4.1.

Anoll, A. P., L. D. Davenport, and R. Lienesch (2024). "Racial Context(s) in American Political Behavior". In: *American Political Science Review*, pp. 1–17. DOI: 10.1017/S0003055424000832.

Banerjee, A. et al. (2015). "A multifaceted program causes lasting progress for the very poor: Evidence from six countries". In: *Science* 348.6236, p. 1260799. DOI: 10.1126/science.1260799.

Bansak, K., E. Paulson, and D. Rothenhausler (2023). *Learning under random distributional shifts*. arXiv: 2306.02948 [stat.ML]. URL: https://arxiv.org/abs/2306.02948.

Ben-Tal, A. et al. (2013). "Robust Solutions of Optimization Problems Affected by Uncertain Probabilities". In: *Management Science* 59.2, pp. 341–357. DOI: 10.1287/mnsc.1120.1641. eprint: https://doi.org/10.1287/mnsc.1120.1641. URL: https://doi.org/10.1287/mnsc.1120.1641.

Bertsimas, D., K. Imai, and M. L. Li (2023). *Distributionally Robust Causal Inference with Observational Data*. arXiv: 2210.08326 [stat.ME]. URL: https://arxiv.org/abs/2210.08326.

Bertsimas, D., M. Johnson, and N. Kallus (2015). "The Power of Optimization Over Randomization in Designing Experiments Involving Small Samples". In: *Operations Research* 63.4, pp. 868–876. DOI: 10.1287/opre.2015.1361.

Bertsimas, D. and M. Sim (2004). "The Price of Robustness". In: *Operations Research* 52.1, pp. 35–53. DOI: 10.1287/opre.1030.0065. eprint: https://doi.org/10.1287/opre.1030.0065. URL: https://doi.org/10.1287/opre.1030.0065.

Blair, G. et al. (2021). "Community policing does not build citizen trust in police or reduce crime in the Global South". In: *Science* 374.6571, eabd3446. DOI: 10.1126/science.abd3446.

Blanchet, J. and K. Murthy (2019). "Quantifying Distributional Model Risk via Optimal Transport". In: *Mathematics of Operations Research* 44.2, pp. 565–600. DOI: 10.1287/moor.2018.0936. eprint: https://doi.org/10.1287/moor.2018.0936. URL: https://doi.org/10.1287/moor.2018.0936.

Blanchet, J., K. Murthy, and V. A. Nguyen (2021). *Statistical Analysis of Wasserstein Distributionally Robust Estimators*. arXiv: 2108.02120 [math.ST]. URL: https://arxiv.org/abs/2108.02120.

Blanchet, J. et al. (2024). *Distributionally Robust Optimization and Robust Statistics*. arXiv: 2401.14655 [stat.ME]. URL: https://arxiv.org/abs/2401.14655.

Bloom, H. S. et al. (2017). "Using Multisite Experiments to Study Cross-Site Variation in Treatment Effects: A Hybrid Approach With Fixed Intercepts and a Random Treatment Coefficient". In: *Journal of Research on Educational Effectiveness* 10.4, pp. 817–842. DOI: 10.1080/19345747.2016.1264518. eprint: https://doi.org/10.1080/19345747.2016.1264518. URL: https://doi.org/10.1080/19345747.2016.1264518.

Bold, T. et al. (2018). "Experimental evidence on scaling up education reforms in Kenya". In: *Journal of Public Economics* 168, pp. 1–20. ISSN: 0047-2727. DOI: https://doi.org/10.1016/j.jpubeco.2018.08.007. URL: https://www.sciencedirect.com/science/article/pii/S0047272718301518.

Bound, J., C. Brown, and N. Mathiowetz (2001). "Chapter 59 - Measurement Error in Survey Data". In: ed. by J. J. Heckman and E. Leamer. Vol. 5. Handbook of Econometrics. Elsevier, pp. 3705–3843. DOI: https://doi.org/10.1016/S1573-4412(01)05012-7. URL: https://www.sciencedirect.com/science/article/pii/S1573441201050127.

Cheeseman, N. and C. Peiffer (2023). "Why efforts to fight corruption can undermine the social contract: Lessons from a survey experiment in Nigeria". In: *Governance* 36.4, pp. 1045–1061. DOI: https://doi.org/10.1111/gove.12720. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/gove.12720. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/gove.12720.

Cohodes, S. and S. Roy (Aug. 2024). "Thirty Years of Charter Schools: What Does Lottery-Based Research Tell Us?" In: *Journal of School Choice* 19.1, pp. 8–49. ISSN: 1558-2167. DOI: 10.1080/15582159.2024.2379644. URL: http://dx.doi.org/10.1080/15582159.2024.2379644.

Crépon, B. et al. (Jan. 2015). "Estimating the Impact of Microcredit on Those Who Take It Up: Evidence from a Randomized Experiment in Morocco". In: *American Economic Journal: Applied Economics* 7.1, pp. 123–50. DOI: 10.1257/app.20130535. URL: https://www.aeaweb.org/articles?id=10.1257/app.20130535.

Daitch, V. et al. (2022). "Underrepresentation of Women in Randomized Controlled Trials: A Systematic Review and Meta-Analysis". In: *Trials* 23.1, p. 1038. DOI: 10.1186/s13063-022-07004-2.

Dakin, R. J. (Nov. 1965). "A tree-search algorithm for mixed integer programming problems". In: *The Computer Journal* 8.3, pp. 250–255. DOI: 10.1093/comjnl/8.3.250.

Dantzig, G. B. (1951). "Application of the Simplex Method to a Transportation Problem". In: *Activity Analysis of Production and Allocation*. Ed. by T. C. Koopmans. New York: J. Wiley & Sons, pp. 359–373.

Dehejia, R., C. Pop-Eleches, and C. Samii (2019). *From Local to Global: External Validity in a Fertility Natural Experiment*. arXiv: 1906.08096 [econ.EM]. URL: https://arxiv.org/abs/1906.08096.

Duchi, J. and H. Namkoong (2020). *Learning Models with Uniform Performance via Distributionally Robust Optimization*. arXiv: 1810.08750 [stat.ML]. URL: https://arxiv.org/abs/1810.08750.

Dunning, T. et al., eds. (2019). *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. Cambridge University Press.

Egami, N. and E. Hartman (Aug. 2021). "Covariate selection for generalizing experimental results: Application to a large-scale development program in Uganda". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 184.4, pp. 1524–1548. ISSN: 1467-985X. DOI: 10.1111/rssa.12734. URL: http://dx.doi.org/10.1111/rssa.12734.

— (2023). "Elements of External Validity: Framework, Design, and Analysis". In: *American Political Science Review* 117.3, pp. 1070–1088. DOI: 10.1017/S0003055422000880.

Egami, N. and D. D. I. Lee (2024). *Designing Multi-Site Studies for External Validity: Site Selection via Synthetic Purposive Sampling*.

Eichler, H.-G. et al. (2011). "Bridging the Efficacy-Effectiveness Gap: A Regulator's Perspective on Addressing Variability of Drug Response". In: *Nature Reviews Drug Discovery* 10.7, pp. 495–506. DOI: 10.1038/nrd3501.

Esfahani, P. M. and D. Kuhn (2017). *Data-driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations*. arXiv: 1505.05116 [math.OC]. URL: https://arxiv.org/abs/1505.05116.

Findley, M. G., K. Kikuta, and M. Denly (May 2021). "External Validity". In: *Annual Review of Political Science* 24.1, pp. 365–393. ISSN: 1545-1577. DOI: 10.1146/annurev-polisci-041719-102556. URL: http://dx.doi.org/10.1146/annurev-polisci-041719-102556.

Gechter, M. et al. (2024). *Selecting Experimental Sites for External Validity*. arXiv: 2405.13241 [econ.GN].

Hassell, H. J. (2021). "Local racial context, campaign messaging, and public political behavior: A congressional campaign field experiment". In: *Electoral Studies* 69, p. 102247. ISSN: 0261-3794. DOI: https://doi.org/10.1016/j.electstud.2020.102247. URL: https://www.sciencedirect.com/science/article/pii/S0261379420301268.

Heckman, J. (1974). "Shadow Prices, Market Wages, and Labor Supply". In: *Econometrica* 42.4, pp. 679–694. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1913937 (visited on 09/09/2026).

Hirano, K. and J. R. Porter (2009). "Asymptotics for Statistical Treatment Rules". In: *Econometrica* 77.5, pp. 1683–1701. DOI: 10.3982/ECTA6630.

Hu, Y. et al. (2024). *Minimax-Regret Sample Selection in Randomized Experiments*. arXiv: 2403.01386 [stat.ME]. URL: https://arxiv.org/abs/2403.01386.

Kallus, N. (2018). "Optimal A Priori Balance in the Design of Controlled Experiments". In: *Journal of the Royal Statistical Society: Series B* 80.1, pp. 85–112. DOI: 10.1111/rssb.12240.

— (2020). "Optimal a priori balance in the design of controlled experiments". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82.4, pp. 1243–1272.

— (2021). "On the Optimality of Randomization in Experimental Design: How to Randomize for Minimax Variance and Design-Based Inference". In: *Journal of the Royal Statistical Society: Series B* 83.2, pp. 404–409. DOI: 10.1111/rssb.12412.

Kantorovich, L. V. (Mar. 2006). "On the Translocation of Masses". In: *Journal of Mathematical Sciences* 133.4, pp. 1381–1382. ISSN: 1573-8795. DOI: 10.1007/s10958-006-0049-2. URL: http://dx.doi.org/10.1007/s10958-006-0049-2.

Kuhn, D. et al. (2024). *Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning*. arXiv: 1908.08729 [stat.ML]. URL: https://arxiv.org/abs/1908.08729.

Land, A. H. and A. G. Doig (1960). "An Automatic Method of Solving Discrete Programming Problems". In: *Econometrica* 28.3, pp. 497–520. DOI: 10.2307/1910129. URL: https://jstor.org.

Levy, D. et al. (2020). *Large-Scale Methods for Distributionally Robust Optimization*. arXiv: 2010.05893 [math.OC]. URL: https://arxiv.org/abs/2010.05893.

Luo, F. and S. Mehrotra (Apr. 2020). "Distributionally robust optimization with decision dependent ambiguity sets". In: *Optimization Letters* 14.8, pp. 2565–2594. ISSN: 1862-4480. DOI: 10.1007/s11590-020-01574-3. URL: http://dx.doi.org/10.1007/s11590-020-01574-3.

Manski, C. F. (2004). "Statistical Treatment Rules for Heterogeneous Populations". In: *Econometrica* 72.4, pp. 1221–1246. DOI: 10.1111/j.1468-0262.2004.00530.x.

Mosteller, F. (1995). "The Tennessee Study of Class Size in the Early School Grades". In: *The Future of Children* 5.2, p. 113. ISSN: 1054-8289. DOI: 10.2307/1602360. URL: http://dx.doi.org/10.2307/1602360.

Munro, E., X. Kuang, and S. Wager (2025). *Treatment Effects in Market Equilibrium*. arXiv: 2109.11647 [econ.EM]. URL: https://arxiv.org/abs/2109.11647.

Oh, S. S. et al. (2015). "Diversity in Clinical and Biomedical Research: A Promise Yet to Be Fulfilled". In: *PLoS Medicine* 12.12, e1001918. DOI: 10.1371/journal.pmed.1001918.

Olea, J. L. M. et al. (2024). *Externally Valid Selection of Experimental Sites via the k-Median Problem*. arXiv: 2408.09187 [econ.EM]. URL: https://arxiv.org/abs/2408.09187.

Olsen, F. et al. (Feb. 2022). "Socioeconomic and geographic differences in ablation of atrial fibrillation in Norway - a national cohort study". In: *BMC Public Health* 22.1. DOI: 10.1186/s12889-022-12628-9. URL: https://doi.org/10.1186/s12889-022-12628-9.

Olsen, R. B. and L. L. Orr (Dec. 2016). "On the "Where" of Social Experiments: Selecting More Representative Samples to Inform Policy". In: *New Directions for Evaluation* 2016.152, pp. 61–71. ISSN: 1534-875X. DOI: 10.1002/ev.20207. URL: http://dx.doi.org/10.1002/ev.20207.

Orr, L. L. et al. (June 2019). "Using the Results from Rigorous Multisite Evaluations to Inform Local Policy Decisions". In: *Journal of Policy Analysis and Management* 38.4, pp. 978–1003. ISSN: 1520-6688. DOI: 10.1002/pam.22154. URL: http://dx.doi.org/10.1002/pam.22154.

Pearl, J. and E. Bareinboim (2011). "Transportability of Causal and Statistical Relations: A Formal Approach". In: *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 540–547. DOI: 10.1109/ICDMW.2011.169.

— (Nov. 2014). "External Validity: From Do-Calculus to Transportability Across Populations". In: *Statistical Science* 29.4. ISSN: 0883-4237. DOI: 10.1214/14-sts486. URL: http://dx.doi.org/10.1214/14-STS486.

Peyré, G. and M. Cuturi (2019). "Computational Optimal Transport". In: *Foundations and Trends in Machine Learning* 11.5-6, pp. 355–607.

Ramamoorthy, A. et al. (2015). "Racial/Ethnic Differences in Drug Disposition and Response: Review of Recently Approved Drugs". In: *Clinical Pharmacology and Therapeutics* 97.3, pp. 263–273. DOI: 10.1002/cpt.61.

Raudenbush, S. W. and D. Schwartz (Mar. 2020). "Randomized Experiments in Education, with Implications for Multilevel Causal Inference". In: *Annual Review of Statistics and Its Application* 7.1, pp. 177–208. ISSN: 2326-831X. DOI: 10.1146/annurev-statistics-031219-041205. URL: http://dx.doi.org/10.1146/annurev-statistics-031219-041205.

Rossouw, J. E. et al. (2002). "Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women: Principal Results From the Women's Health Initiative Randomized Controlled Trial". In: *JAMA* 288.3, pp. 321–333. DOI: 10.1001/jama.288.3.321.

Rothenhausler, D. and P. Bühlmann (2023). *Distributionally robust and generalizable inference.* arXiv: 2209.09352 [stat.ME]. URL: https://arxiv.org/abs/2209.09352.

Rudolph, K. E. and I. Díaz (Feb. 2021). "Efficiently transporting causal direct and indirect effects to new populations under intermediate confounding and with multiple mediators". In: *Biostatistics* 23.3, pp. 789–806. ISSN: 1468-4357. DOI: 10.1093/biostatistics/kxaa057. URL: http://dx.doi.org/10.1093/biostatistics/kxaa057.

Rudolph, K. E. et al. (2024). *Improving efficiency in transporting average treatment effects.* arXiv: 2304.00117 [stat.ME]. URL: https://arxiv.org/abs/2304.00117.

Saville, B. R. et al. (Aug. 2022). "The Bayesian Time Machine: Accounting for temporal drift in multi-arm platform trials". In: *Clinical Trials* 19.5, pp. 490–501. ISSN: 1740-7753. DOI: 10.1177/17407745221112013. URL: http://dx.doi.org/10.1177/17407745221112013.

Shadish, W., T. Cook, and D. Campbell (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference.* Experimental and Quasi-experimental Designs for Generalized Causal Inference v. 1. Houghton Mifflin. ISBN: 9780395615560. URL: https://books.google.com/books?id=o7jaAAAAMAAJ.

Stoye, J. (2009). "Minimax Regret Treatment Choice with Finite Samples". In: *Journal of Econometrics* 151.1, pp. 70–81. DOI: 10.1016/j.jeconom.2009.02.013.

Tan, Y. Y. et al. (Oct. 2022). "Comparing clinical trial population representativeness to real-world populations: an external validity analysis encompassing 43895 trials and 5685738 individuals across 989 unique drugs and 286 conditions in England". In: *The Lancet Healthy Longevity* 3.10, e674–e689. ISSN: 2666-7568. DOI: 10.1016/s2666-7568(22)00186-6. URL: http://dx.doi.org/10.1016/S2666-7568(22)00186-6.

Tibshirani, R. J. et al. (2020). *Conformal Prediction Under Covariate Shift*. arXiv: 1904.06019 [stat.ME]. URL: https://arxiv.org/abs/1904.06019.

Tipton, E. (2013a). "Improving Generalizations From Experiments Using Propensity Score Subclassification: Assumptions, Properties, and Contexts". In: *Journal of Educational and Behavioral Statistics* 38.3, pp. 239–266. DOI: 10.3102/1076998612441947. eprint: https://doi.org/10.3102/1076998612441947. URL: https://doi.org/10.3102/1076998612441947.

— (Apr. 2013b). "Stratified Sampling Using Cluster Analysis: A Sample Selection Strategy for Improved Generalizations From Experiments". In: *Evaluation Review* 37.2, pp. 109–139. ISSN: 1552-3926. DOI: 10.1177/0193841x13516324. URL: http://dx.doi.org/10.1177/0193841X13516324.

Van Spall, H. G. C. et al. (2007). "Eligibility Criteria of Randomized Controlled Trials Published in High-Impact General Medical Journals: A Systematic Sampling Review". In: *JAMA* 297.11, pp. 1233–1240. DOI: 10.1001/jama.297.11.1233.

Villani, C. (2008). *Optimal Transport: Old and New*. Vol. 338. Grundlehren der Mathematischen Wissenschaften. Berlin: Springer.

Wald, A. (1950). *Statistical Decision Functions*. New York: John Wiley & Sons.

WHO Solidarity Trial Consortium et al. (2021). "Repurposed Antiviral Drugs for Covid-19 — Interim WHO Solidarity Trial Results". In: *New England Journal of Medicine* 384.6, pp. 497–511. DOI: 10.1056/NEJMoa2023184.

Zhang, Y., M. Huang, and K. Imai (2024). *Minimax Regret Estimation for Generalizing Heterogeneous Treatment Effects with Multisite Data*. arXiv: 2412.11136 [stat.ME]. URL: https://arxiv.org/abs/2412.11136.

Zulman, D. M. et al. (2011). "Examining the Evidence: A Systematic Review of the Inclusion and Analysis of Older Adults in Randomized Controlled Trials". In: *Journal of General Internal Medicine* 26.7, pp. 783–790. DOI: 10.1007/s11606-010-1629-x.

A Proofs

Preliminaries

We observe $|P|$ candidate sites indexed by s, and want to select a subset of size K. Each site has a population size of N_s. We write $p_s = N_s / \sum_r N_r$ for the share of all individuals who are in site s, where $p_s > 0$ for each s and $\sum_s p_s = 1$. In each site, units are sampled into the experiment according to a sampling indicator $W_s \sim \Phi_s$ an experiment is conducted, and individuals are randomized into treatmen treatment Z_s according to a design Ψ_s. Each site has observed features $X_s \in \mathbb{R}^d$, and unobserved features $U_s \in \mathbb{R}^k$.

$P_{X,U} = P$ is the joint population distribution on $\mathbb{R}^d \times \mathbb{R}^k$ with mass p_s at the point (X_s, U_s), for each site s. P_X, the observed covariate distribution, places mass p_s at each site-level covariate vector X_s. P_U is the distribution of unobservables, and $P_{U|X}$ is the population distribution of unobservables with observables fixed at $X = x$. $S_{X,U}$, S_X, S_U, $S_{U|X}$ are defined analogously. For a deployment population Q, $Q_{X,U}$, Q_X, Q_U, $Q_{U|X}$ are defined in the same way.

Individual i in site s has treatment effect τ_{is}. We write $\tau_s = \frac{1}{N_s}\sum_{i\in s}\tau_{is}$ for the average treatment effect among the N_s units in site s.

Let $\tau : \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}$ be the treatment effect function, so that $\tau_s = \tau(X_s, U_s)$. Then, for a distribution F over (X, U), define:

$$\mathrm{ATE}(F) \;=\; \int \tau(X, U)\, dF(X, U).$$

Evaluated over the observed candidate population:

$$\mathrm{ATE}(P) = \sum_s p_s \tau_s$$

This is the average effect over individuals in the population. This is the PATE, but it is helpful to make the dependence on the set P explicit.

Given a selected subset S, define the average treatment effect over the selected sites:

$$\mathrm{ATE}(S) = \frac{1}{K}\sum_{s\in S}\tau_s$$

This is the Sample Average Treatment Effect (SATE) evaluated on the selected subset S. Since S is the choice variable, it is again helpful to write this as a function that ranges over S.

For a hypothetical deployment population Q, we likewise define the estimand of interest:

$$\mathrm{ATE}(Q) = \int \tau(X, U)\, dQ(X, U)$$

Let ω_s be a vector of weights on the selected sites, and $\hat{\tau}_s$ the plug-in estimate of τ_s from the experiment run in site s. Then:

$$\hat{\tau}(S) \;=\; \sum_{s\in S}\omega_s\, \hat{\tau}_s.$$

Remark 2 (Balanced transport)**.** *Throughout, $\omega_s = 1/K$, so that $\hat{\tau}(S)$ is the average of the site-level estimates and $\mathbb{E}[\hat{\tau}(S)] = \mathrm{ATE}(S)$. Each selected site stands in for an equal share of the population.*

Recall the definition of the 1-Wasserstein distance above. For distributions μ, ν on $\mathbb{R}^d$ (or on $\mathbb{R}^k$) with finite first moments, let $\Pi(\mu, \nu)$ be the set of couplings of μ and ν, and

$$W_1(\mu, \nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \int \|x - x'\|\, d\gamma(x, x').$$

For the joint space (X, U) with support $\mathbb{R}^d \times \mathbb{R}^k$ define the distance metric:

$$d\big((x,u),(x',u')\big) = \|x - x'\| + \|u - u'\|,$$

and, for joint distribution F, G on $\mathbb{R}^d \times \mathbb{R}^k$ with finite first moments,

$$W_1^{X,U}(F,G) = \inf_{\gamma \in \Pi(F,G)} \int d\big((x,u),(x',u')\big)\, d\gamma\big((x,u),(x',u')\big).$$

This is the 1-Wasserstein distance between two distributions F and G. This represents the distance between F and G under the optimal transport plan from F to G in (X, U)-space. Every distribution in this appendix is supported on finitely many sites, so each integral is a finite sum and each transport plan is a matrix.

A.1 Proof of Lemma 2.1

This is a familiar result.

Proof.

$$\begin{aligned}
\mathbb{E}\big[\widehat{ATE}(S)\big] &= \mathbb{E}\left[\frac{1}{K}\sum_{s\in S}\hat{\tau}_s\right] \\
&= \frac{1}{K}\sum_{s\in S}\mathbb{E}[\hat{\tau}_s] \\
&= \frac{1}{K}\sum_{s\in S}\mathbb{E}_{Z_s\sim\Psi_s,\, W_s\sim\Phi_s}[\hat{\tau}_s] \\
&= \frac{1}{K}\sum_{s\in S}\tau(X_s, U_s) \\
&= \int \tau(x,u)\, dS_{X,U}(x,u) \\
&= ATE(S)
\end{aligned}$$

□

A.2 Proof of Lemma 2.2

Proof.

$$\begin{aligned}
\mathrm{MSE}_Q(S) &= \mathbb{E}\Big[\big(\mathrm{ATE}(Q) - \widehat{\mathrm{ATE}}(S)\big)^2\Big] \\
&= \mathbb{E}\Big[\big([\mathrm{ATE}(Q) - \mathrm{ATE}(S)] - [\widehat{\mathrm{ATE}}(S) - \mathrm{ATE}(S)]\big)^2\Big] \\
&= \big(\mathrm{ATE}(Q) - \mathrm{ATE}(S)\big)^2 - 2\big(\mathrm{ATE}(Q) - \mathrm{ATE}(S)\big)\,\mathbb{E}\big[\widehat{\mathrm{ATE}}(S) - \mathrm{ATE}(S)\big] \\
&\qquad + \mathbb{E}\Big[\big(\widehat{\mathrm{ATE}}(S) - \mathrm{ATE}(S)\big)^2\Big] \\
&= \big(\mathrm{ATE}(Q) - \mathrm{ATE}(S)\big)^2 + \mathbb{E}\Big[\big(\widehat{\mathrm{ATE}}(S) - \mathrm{ATE}(S)\big)^2\Big] \qquad \text{[By Lemma 2.1]} \\
&= \big(\mathrm{ATE}(Q) - \mathrm{ATE}(S)\big)^2 + \sigma_S^2
\end{aligned}$$

□

A.3 A Lipschitz–transport bound

Lemma A.1 (Lipschitz–transport bound)**.** *Let τ be an L-Lipschitz function. Then, under Assumption 2.1:*

$$\Big|\int \tau\, dF - \int \tau\, dG\Big| \;\leq\; L\, W_1^{X,U}(F,G).$$

Proof. Consider a transport plan $\gamma \in \Gamma(F,G)$.

$$\begin{aligned}
\int \tau\, dF - \int \tau\, dG &= \int \big[\tau(x,u) - \tau(x',u')\big]\, d\gamma\big((x,u),(x',u')\big) \\
\Big|\int \tau\, dF - \int \tau\, dG\Big| &\leq \int \big|\tau(x,u) - \tau(x',u')\big|\, d\gamma \\
&\leq L \int \big(\|x-x'\| + \|u-u'\|\big)\, d\gamma && \text{[By Assumption 2.1]} \\
&= L \int d\big((x,u),(x',u')\big)\, d\gamma
\end{aligned}$$

This establishes that:

$$\Big|\int \tau\, dF - \int \tau\, dG\Big| \leq L \int d\big((x,u),(x',u')\big)\, d\gamma$$

But since γ was an arbitrary transport plan, it follows that this inequality holds under the *optimal* transport plan:

$$\Big|\int \tau\, dF - \int \tau\, dG\Big| \leq \inf_{\gamma} L \int d\big((x,u),(x',u')\big)\, d\gamma$$

So that:

$$\Big|\int \tau\, dF - \int \tau\, dG\Big| \leq L W_1^{X,U}(F,G)$$

□

In words, this establishes that the absolute distance between a Lipschitz function $\tau(X,U)$ evaluate on two distributions F and G is upper bounded by a Lipschitz constant times the cost of the optimal transport plan from F to G.

A.4 Proof of Theorem 2.3

Proof.

$$\begin{aligned}
\mathrm{MSE}_Q(S) &= \big(\mathrm{ATE}(Q) - \mathrm{ATE}(S)\big)^2 + \sigma_S^2 && \text{[By Lemma 2.2]} \\
\mathrm{ATE}(Q) - \mathrm{ATE}(S) &= \int \tau(x)\, dQ_X(x) - \int \tau(x)\, dS_X(x) \\
\big|\mathrm{ATE}(Q) - \mathrm{ATE}(S)\big| &\leq L\, W_1(Q_X, S_X) && \text{[By Lemma A.1]} \\
\mathrm{MSE}_Q(S) &\leq L^2\, W_1(Q_X, S_X)^2 + \sigma_S^2
\end{aligned}$$

□

A.5 Bound on unobservables

Lemma A.2 (Bounding the unobservables)**.** *Let $\gamma^* \in \Pi(S_X, Q_X)$ be an optimal coupling of S_X and Q_X, with first coordinate $x \sim S_X$ and second coordinate $x' \sim Q_X$. Define*

$$\eta_{\gamma^*}(S,Q) \;=\; \int W_1\big(S_{U|X=x},\, Q_{U|X=x'}\big)\, d\gamma^*(x,x'),$$

the average distance between the conditional distributions of the unobservables, given the optimal transport plan γ^*. Then:

$$W_1^{X,U}(S_{X,U}, Q_{X,U}) \;\le\; W_1(S_X, Q_X) + \eta_{\gamma^*}(S,Q)$$

Proof. For each pair (x,x') in the support of γ^*, let $\kappa_{x,x'}$ be a transport plan from $Q_{U|X=x'}$ to $S_{U|X=x}$ that attains $W_1(S_{U|X=x}, Q_{U|X=x'})$.

Define a measure $\hat{\Gamma}$ on $(\mathbb{R}^d \times \mathbb{R}^k) \times (\mathbb{R}^d \times \mathbb{R}^k)$ by

$$\hat{\Gamma}(dx, du, dx', du') = \gamma^*(dx, dx')\, \kappa_{x,x'}(du, du')$$

First, we show that $\hat{\Gamma}$ is a coupling of $Q_{X,U}$ and $S_{X,U}$. We first show that integrating over $\hat{\Gamma}$ yields these marginals.

First, integrating out (x,u), we get the source distribution:

$$\begin{aligned}
\int_{x,u} \hat{\Gamma}(dx, du, dx', du') &= \int_x \gamma^*(dx, dx') \int_u \kappa_{x,x'}(du, du') \\
&= \int_x \gamma^*(dx, dx')\, Q_{U|X=x'}(du') \\
&= Q_{U|X=x'}(du') \int_x \gamma^*(dx, dx') \\
&= Q_{U|X=x'}(du')\, Q_X(dx') \\
&= Q_{X,U}(dx', du')
\end{aligned}$$

Second, integrating out (x',u'), we get the target distribution:

$$\begin{aligned}
\int_{x',u'} \hat{\Gamma}(dx, du, dx', du') &= \int_{x'} \gamma^*(dx, dx') \int_{u'} \kappa_{x,x'}(du, du') \\
&= \int_{x'} \gamma^*(dx, dx')\, S_{U|X=x}(du) \\
&= S_X(dx)\, S_{U|X=x}(du) \\
&= S_{X,U}(dx, du)
\end{aligned}$$

So $\hat{\Gamma}$ has first marginal $S_{X,U}$ and second marginal $Q_{X,U}$: it is a coupling of the two, and so a transport plan between them.

Next, we show that the cost of the transport plan $\hat{\Gamma}$ is the cost of the optimal transport plan from Q_X to S_X plus the distance from the unobservables of Q to the unobservables of S under the optimal transport plan from Q_X to S_X.

$$
\begin{aligned}
\int d\, d\hat{\Gamma} &= \int \left(\|x-x'\| + \|u-u'\|\right) \gamma^*(dx,dx')\, \kappa_{x,x'}(du,du') \\
&= \int \|x-x'\|\, d\gamma^*(x,x') + \int \Big[\int \|u-u'\|\, d\kappa_{x,x'}(u,u') \Big] d\gamma^*(x,x') \\
&= W_1(S_X,Q_X) + \int W_1\big(S_{U|X=x}, Q_{U|X=x'}\big)\, d\gamma^*(x,x') \\
&= W_1(S_X,Q_X) + \eta(S_U,Q_U)
\end{aligned}
$$

By definition of the Wasserstein distance under the joint metric d,

$$
W_1^{X,U}\big(S_{X,U}, Q_{X,U}\big) \;=\; \inf_{\Gamma \in \Pi(S_{X,U},Q_{X,U})} \int d\big((x,u),(x',u')\big)\, d\Gamma(x,u,x',u'),
$$

where $\Pi(S_{X,U}, Q_{X,U})$ is the set of all couplings of $S_{X,U}$ and $Q_{X,U}$. Therefore

$$
W_1^{X,U}\big(S_{X,U}, Q_{X,U}\big) = \inf_{\Gamma \in \Pi(S_{X,U},Q_{X,U})} \int d\, d\Gamma \le \int d\, d\hat{\Gamma}
$$

So that:

$$
W_1^{X,U}\big(S_{X,U}, Q_{X,U}\big) \le W_1(S_X,Q_X) + \eta(S_U,Q_U)
$$

□

A.6 Proof of Theorem 2.4

Proof.

$$
\begin{aligned}
\mathrm{MSE}_Q(S) &= \big(\mathrm{ATE}(Q) - \mathrm{ATE}(S)\big)^2 + \sigma_S^2 && \text{[By Lemma 2.2]} \\
\big|ATE(Q) - ATE(S)\big| &= \Big| \int \tau\, dQ_{X,U} - \int \tau\, dS_{X,U} \Big| && \text{[By Definition 2.1]} \\
&\le L\, W_1^{X,U}(S_{X,U}, Q_{X,U}) && \text{[By Lemma A.1]} \\
&\le L\left[W_1(S_X,Q_X) + \eta(S_U,Q_U)\right] && \text{[By Lemma A.2]} \\
\mathrm{MSE}_Q(S) &\le L^2\big[W_1(S_X,Q_X) + \eta(S_U,Q_U)\big]^2 + \sigma_S^2 \quad \square
\end{aligned}
$$

A.7 Proof of Theorem 3.1

Recall that $\mathcal{Q}_h$ is the restriction of Θ_ρ to a grid. Define two optimal values, V, the optimal value over a directional shift family, and V_h, the optimal value over a grid with mesh size h:

$$
V = \min_{|S|=K} \sup_{\theta \in \Theta_\rho} W_1(Q_\theta, S_X), \qquad V_h = \min_{|S|=K} \max_{Q \in \mathcal{Q}_h} W_1(Q, S_X).
$$

Lemma A.3 (Grid approximation error)**.** *For every selection* S*,*

$$\max_{Q \in \mathcal{Q}_h} W_1(Q, S_X) \;\leq\; \sup_{\theta \in \Theta_\rho} W_1(Q_\theta, S_X) \;\leq\; \max_{Q \in \mathcal{Q}_h} W_1(Q, S_X) + c\, h,$$

and consequently $V_h \leq V \leq V_h + c\, h$*.*

Proof. We show that approximating the parameter space by points on a grid adds error ch to the upper bound minimized by Algorithm 1.

For any S and any θ, θ':

$$\begin{aligned} W_1(Q_\theta, S_X) \;&\leq\; W_1(Q_{\theta'}, S_X) + W_1(Q_\theta, Q_{\theta'}) && \text{[Triangle inequality]} \\ &\leq\; W_1(Q_{\theta'}, S_X) + c\, \|\theta - \theta'\| && \text{[Lipschitz]} \end{aligned}$$

Every $\theta \in \Theta_\rho$ has a grid point θ' within h, and $\mathcal{Q}_h \subseteq \{Q_\theta : \theta \in \Theta_\rho\}$, so

$$\max_{Q \in \mathcal{Q}_h} W_1(Q, S_X) \;\leq\; \sup_{\theta \in \Theta_\rho} W_1(Q_\theta, S_X) \;\leq\; \max_{Q \in \mathcal{Q}_h} W_1(Q, S_X) + c\, h \qquad \forall S.$$

Minimizing over S: $V_h \leq V \leq V_h + c\, h$. ☐

Lemma A.4 (The bounds bracket the optimum)**.** *At each iteration* t*,*

$$\mathrm{LB}^{(t)} \;\leq\; V_h \;\leq\; \mathrm{UB}^{(t)}.$$

Proof. Recall that at each t the algorithm stores a proposed site selection $\tilde{S}^{(t)}$, and a collection $\tilde{\mathcal{Q}}^{(t)} \subseteq \mathcal{Q}_h$ of adversarial shifts observed until t. From these it computes an upper bound and a lower bound on the worst-case site selection.

First, the upper bound. The adversary evaluates the current site selection $\tilde{S}^{(t)}$ against every scenario on the grid and returns the worst observed loss.

$$\mathrm{UB}^{(t)} = \max_{Q \in \mathcal{Q}_h} W_1\big(Q, \tilde{S}^{(t)}_X\big).$$

V_h is the same worst case minimized over *all* selections, so it is smaller than the worst-case chosen by the adversary:

$$V_h = \min_{|S|=K} \max_{Q \in \mathcal{Q}_h} W_1(Q, S_X) \;\leq\; \max_{Q \in \mathcal{Q}_h} W_1\big(Q, \tilde{S}^{(t)}_X\big) = \mathrm{UB}^{(t)}.$$

Next, we consider the lower bound. The planner solves a mixed-integer program whose value is the best worst case over the scenarios found so far:

$$\mathrm{LB}^{(t)} = \min_{|S|=K} \max_{Q \in \tilde{\mathcal{Q}}^{(t)}} W_1(Q, S_X).$$

Now compare with V_h. For any fixed selection S, the scenarios in $\tilde{\mathcal{Q}}^{(t)}$ are a subset of the scenarios in $\mathcal{Q}_h$, so:

$$\max_{Q \in \tilde{\mathcal{Q}}^{(t)}} W_1(Q, S_X) \;\leq\; \max_{Q \in \mathcal{Q}_h} W_1(Q, S_X) \qquad \text{for every } S$$

And so, in particular:

$$\min_{|S|=K} \max_{Q \in \tilde{\mathcal{Q}}^{(t)}} W_1(Q, S_X) \ \leq \ \min_{|S|=K} \max_{Q \in \mathcal{Q}_h} W_1(Q, S_X)$$

Where, plugging in definitions, we have:

$$\mathrm{LB}^{(t)} = \min_{|S|=K} \max_{Q \in \tilde{\mathcal{Q}}^{(t)}} W_1(Q, S_X) \ \leq \ \min_{|S|=K} \max_{Q \in \mathcal{Q}_h} W_1(Q, S_X) = V_h.$$

$\square$

Lemma A.5 (Algorithm stopping rule)**.** *Algorithm 1 stops within $|\mathcal{Q}_h|$ iterations, and when it stops, $\min_{u<t} \mathrm{UB}^{(u)} - \mathrm{LB}^{(t)} \leq \epsilon$.*

Proof. Recall that at each iteration $t+1$, the worst-case perturbation $\tilde{Q}^{(t+1)}$ is added to the collection, $\tilde{\mathcal{Q}}^{(t+1)} = \tilde{\mathcal{Q}}^{(t)} \cup \{\tilde{Q}^{(t+1)}\}$.

The planner then solves the master problem over the enlarged collection, and the value it attains is a new lower bound over this larger set:

$$\mathrm{LB}^{(t+1)} = \max_{Q \in \tilde{\mathcal{Q}}^{(t+1)}} W_1\big(Q, \tilde{S}_X^{(t+1)}\big) = \min_{|S|=K} \max_{Q \in \tilde{\mathcal{Q}}^{(t+1)}} W_1(Q, S_X).$$

The upper bound at iteration t is the distance from the current selection to the worst scenario the adversary found for it, searching the whole grid:

$$\mathrm{UB}^{(t)} = W_1\big(\tilde{Q}^{(t+1)}, \tilde{S}_X^{(t)}\big) = \max_{Q \in \mathcal{Q}_h} W_1\big(Q, \tilde{S}_X^{(t)}\big).$$

The stopping rule is that the best upper bound seen so far is within ϵ of the current lower bound:

$$\min_{u \leq t} \mathrm{UB}^{(u)} - \mathrm{LB}^{(t+1)} \ \leq \ \epsilon.$$

We show that this condition must be met within $|\mathcal{Q}_h|$ iterations for any ϵ. The basic idea is that the gap is exactly zero whenever the adversary's best response to the planner is an adversarial shift that the planner has already observed.

Suppose that $\tilde{Q}^{(t+1)} \in \tilde{\mathcal{Q}}^{(t)}$. Then, the planner solves the same master problem as before, so that:

$$\mathrm{LB}^{(t+1)} = \mathrm{LB}^{(t)}$$

Now compare the two bounds. Both are worst cases of the same selection $\tilde{S}^{(t)}$, but over different sets of scenarios: the upper bound over the whole grid, the lower bound over the collection found so far:

$$\mathrm{UB}^{(t)} = \max_{Q \in \mathcal{Q}_h} W_1\big(Q, \tilde{S}_X^{(t)}\big), \qquad \mathrm{LB}^{(t)} = \max_{Q \in \tilde{\mathcal{Q}}^{(t)}} W_1\big(Q, \tilde{S}_X^{(t)}\big).$$

The scenario that attains the first maximum is $\tilde{Q}^{(t+1)}$. If it already lies in $\tilde{\mathcal{Q}}^{(t)}$, then it is also the lower bound, and hence:

$$\mathrm{UB}^{(t)} = \mathrm{LB}^{(t)}.$$

Which implies that $\min_{u \le t} \mathrm{UB}^{(u)} - \mathrm{LB}^{(t+1)} \le \mathrm{UB}^{(t)} - \mathrm{LB}^{(t)} = 0 \le \epsilon$, and hence the algorithm stops.

This is guaranteed to occur, since the parameter space contains finitely many points. Otherwise $|\tilde{\mathcal{Q}}^{(t+1)}| = |\tilde{\mathcal{Q}}^{(t)}| + 1$. Since $|\tilde{\mathcal{Q}}^{(0)}| = 1$ and $\tilde{\mathcal{Q}}^{(t)} \subseteq \mathcal{Q}_h$ for all t, this happens at most $|\mathcal{Q}_h| - 1$ times: the algorithm stops within $|\mathcal{Q}_h|$ iterations. □

Proof of Theorem 3.1. Termination is characterized by Lemma A.5. Let t be the stopping iteration and $u^\star = \arg\min_{u<t} \mathrm{UB}^{(u)}$, so $S^*_{\mathrm{DRO}} = \tilde{S}^{(u^\star)}$. Then:

$$
\begin{aligned}
\sup_{\theta \in \Theta_\rho} W_1\big(Q_\theta, S^*_{\mathrm{DRO},X}\big) &\le \max_{Q \in \mathcal{Q}_h} W_1\big(Q, S^*_{\mathrm{DRO},X}\big) + c\,h && \text{[Lemma A.3]} \\
&= \mathrm{UB}^{(u^\star)} + c\,h && \text{[definition of UB]} \\
&\le \mathrm{LB}^{(t)} + \epsilon + c\,h && \text{[Lemma A.5]} \\
&\le V_h + \epsilon + c\,h && \text{[Lemma A.4]} \\
&\le V + \epsilon + c\,h && \text{[Lemma A.3]}
\end{aligned}
$$

□