Experimental Site Selection Under Distribution Shift via Optimal Transport and Wasserstein DRO

Adam Bouyamourn*

October 23, 2025

Abstract

How should researchers select experimental sites when the deployment population may differ from observed data? I formulate the problem of experimental site selection as an *optimal transport problem*, developing methods to minimize downstream estimation error by choosing sites that minimize Wasserstein distances between population and sample covariate distributions. I develop new theoretical upper bounds on PATE and CATE estimation errors, and show that these different objectives lead to different site selection strategies. I extend this approach by using Wasserstein Distributionally Robust Optimization to guard against distribution shift when observed sites may not represent the target population, and develop a novel, data-driven procedure for uncertainty radius selection. Simulation evidence, and a reanalysis of a randomized microcredit experiment in Morocco (Crépon et al.), show that these methods outperform random and stratified sampling of sites, and alternative optimization methods i) for moderate-to-large size problem instances ii) when covariates are moderately informative about treatment effects, and iii) under induced distribution shift.

Keywords: Site Selection, Experimental Design, External Validity, Optimal Transport, Wasserstein Distributionally-Robust Optimization, Causal Inference, Mixed Integer Linear Programming

Contents

[*] Postdoctoral Research Associate in Politics, Statistics and Machine Learning, Department of Politics, Princeton University. I am grateful to Kirk Bansak, Eli Ben-Michael, Peng Ding, Thad Dunning, Naoki Egami, Avi Feller, Adam Glynn, Erin Hartman, Dean Knox, Sam Pimentel, Tara Slough, Anton Strezhnev and participants at EITM 2024, PolMeth 2025, and the Berkeley Methods Workshop for helpful feedback. An R package to implement the approach, `wassersite`, will be made available on publication.

Notation

Symbol	Definition	Notes
Sets and Populations		
$\mathscr{P}$	Universe of all potential experimental sites	Target population
P	Observed subpopulation, $P \subseteq \mathscr{P}$	
S	Selected experimental sites, $S \subset P$	
N	Number of candidate sites, $N = \|P\|$	Population size
K	Maximum sites to select (budget constraint)	
d	Dimension of covariate space	
Covariates and Treatment Effects		
X, U	Observed, unobserved covariates	$X \in \mathbb{R}^d$
x_i	Covariate vector for site i	
P_X, S_X	Empirical covariate distributions	$\frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$, $\frac{1}{K}\sum_{j \in S} \delta_{x_j}$
$Y(1), Y(0)$	Potential outcomes under treatment/control	
$\tau, \tau(x)$	Individual treatment effect, CATE	
τ^P	Population Average Treatment Effect	$\mathbb{E}_{\mathscr{P}}[\tau]$
$\hat{\tau}^S, \hat{\tau}^S(x)$	Sample estimates of PATE, CATE	
Error Measures		
MSE_{PATE}	Mean squared error of PATE estimate	$\mathbb{E}[(\tau^P - \hat{\tau}^S)^2]$
$PEHE$	$\int_X [\tau^P(x) - \hat{\tau}^S(x)]^2 dx$	
L	Lipschitz constant	
η_1, η_2	Sensitivity parameters	
σ_S^2	Irreducible estimation error	
Optimal Transport		
$W_p(P, Q)$	p-Wasserstein distance between P, Q	
π_{ij}, π^*	Transport plan entries, optimal plan	
$c(x, y)$	Transport cost function	$c(x, y) = \\|x - y\\|^p$
δ_{x_i}	Dirac measure at x_i	
Distributionally Robust Optimization		
ρ	Robustness radius	Uncertainty budget
$\mathcal{B}(P_X, \rho)$	Wasserstein ambiguity set	$\{Q : W_p(Q, P_X) \leq \rho\}$
$S^{(t)}$	Site selection at iteration t	
$Q^{(t)}$	Adversarial distribution at iteration t	
$\mathcal{Q}^{(t)}$	Set of adversarial scenarios up to iteration t	
ϵ	Convergence tolerance	
Optimization Variables		
s_i	Binary site selection indicator	$s_i = \mathbb{I}\{i \in S\}$
π_{jk}	MILP transport variables	Population site j to selected site k
μ_k	Adversarial distribution weights	
α_{ik}, β_{kj}	Two-stage transport variables	

Continued on next page

Symbol	Definition	Notes
Simulation and Evaluation		
η	Unmeasured confounding (simulation)	Distinct from η_1, η_2
R^2	Treatment effect variance explained	
$J(S_1, S_2)$	Jaccard similarity	$\frac{\lvert S_1 \cap S_2 \rvert}{\lvert S_1 \cup S_2 \rvert}$
Other		
Z_i	Treatment assignment indicator	
$\mathcal{C}, \mathbf{r}$	Partition, cluster representatives	
$\mathrm{Lip}_1(\mathbb{R}^d)$	Class of 1-Lipschitz functions	

1 Introduction

1.1 Learning from Multi-Site Experimental Studies

Multi-site experimental studies have become central to causal inference across a number of disciplines, as they allow researchers to generate transportable, externally valid estimates of treatment effects that can inform policy-making, theory development, and testing (Bloom et al. 2017; Dunning et al. 2019a).

Across political science, economics, public health, and climate science, multi-site experimental studies are supported by international funding bodies with a view to providing insights that both generalize across multiple contexts, and support cumulative learning about a phenomenon of interest (Dunning et al. 2019a).[1]

In each of these multi-site experimental designs, researchers faced the following problem: given a finite budget and a universe of potential experimental sites, where should they actually conduct an experiment, given their downstream objective of running a valid causal inference experiment that has the smallest possible error estimate?

Further, what should researchers do when deployment populations differ from their observed populations? How should they take into account their limited information about target populations? And how should their decision change when they care about heterogeneity, and ensuring that diverse populations are included in the study sample?

External validity concerns whether findings from an experimental study can be generalized beyond the specific sample, setting, and time period in which the study was conducted (Shadish et al. 2002; Findley et al. 2021; Egami et al. 2023). One central goal within the external validity literature is to develop methods for transporting causal estimates from experimental samples to target populations of policy interest, ensuring that conclusions drawn from studies remain valid when applied to new contexts (Pearl et al. 2011; Pearl et al. 2014; Rudolph et al. 2021; Egami et al. 2021; Rudolph et al. 2024).

Transporting results from one study context is difficult, because it requires us to make assumptions about lack of change between source and target context. In practice, we are likely to encounter **distribution shift**: systematic differences in the distribution of observed covariates between the experimental sample and the target population (Rothenhäusler et al. 2023; Jin et al. 2024). When the covariate distributions differ substantially, estimates derived from the experimental sample may not accurately represent treatment effects in the target population, undermining the external validity of findings.

[1] These notions are fundamentally related: a finding is replicable only if it is externally valid (Jin et al. 2023).

For the experimental planner, distribution shift can take on a number of concrete forms. Feasible experimental sites may differ systematically from the populations on which the researcher wishes to experiment (Allcott 2015). Population characteristics may change in the time period between study planning and implementation (Saville et al. 2022; Bansak et al. 2023). Observed covariates may imperfectly capture true population characteristics, and minority groups may be systematically underrepresented in selected experimental units (Tan et al. 2022; Hu et al. 2024). In political science, differences in institutional quality (Bold et al. 2018), trust in institutions (Cheeseman et al. 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.

Suppose, for instance, we observe rural and urban sites, and use census data at a particular time to estimate population totals in both locations. Census data may suffer from selection bias: it may systematically undercount population totals in hard-to-reach, rural areas. There may also be temporal drift: in-migration from rural to urban areas may have shifted the composition of sites. In each case, the data we *observe* (historic census data) may not be representative of the *deployment population*: the actual rural-urban mix.

How should researchers account for the routine fact that the data they have collected may not accurately represent the population they are in fact interested in (Taori et al. 2020; Rothenhäusler et al. 2023; Bansak et al. 2023; Cai et al. 2023; Jin et al. 2024)?

A goal of recent research in site selection is to choose experimental locations that are, in relevant sense, robust to distribution shift, or designed with external validity in mind (Gechter et al. 2024; Egami et al. 2024; Olea et al. 2024). There a number of different ways we might want to formalize this idea in practice, using different statistical and theoretical tools.

Weighting-based methods aim to improve the external validity of an estimate by reweighting source data so that it more closely matches a prespecified target population (Egami et al. 2021; Huang et al. 2023; Zhang et al. 2024). These methods require that the analyst has a specific transport target in mind and has collected covariate data from the target location. They also require assumptions about the stability of the mapping from source to target: that there is a unique map from source to target, which is learnable.

In contrast, **Distributionally Robust Optimization** (DRO) is 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 et al. 2017; Kuhn et al. 2024; Blanchet et al. 2021; Blanchet et al. 2019a; Blanchet et al. 2024; Duchi et al. 2020; Levy et al. 2020). DRO methods approach the problem of uncertainty guarantee that a 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 et al. 2020; Duchi et al. 2020). Instead of asking, under what assumptions can we transport a valid conclusion from context A to context B, these approaches ask, what solution would we pick if we wanted it to still hold for any context that was ρ-close to the context we actually saw?

These methods build on the *optimal transport* literature, which is an elegant body of applied mathematics that studies the abstract problem of moving (probability) mass from one location to another (Villani et al. 2003; Villani 2008; Bertsimas et al. 2023).

I contrast these methods with sampling-based site selection methods. Throughout this paper, I distinguish between three selection approaches: (1) simple random sampling without stratification, (2) stratified random sampling that randomizes within predefined strata, and (3) optimization-based selection using covariate information. The benefit of simple random sampling is that it does not require prior information about sites, and has optimality and robustness guarantees in general settings. Random sampling is minimax optimal when the analyst has no prior information about experimental units (Kallus 2020), and when the analyst knows the true underlying treatment response only with some error (Wu 1981). Stratification represents a middle ground, using covariates to define strata but maintaining randomization within them.

Optimization methods instead seek to exploit prior information about experimental sites in the form of covariate information. The goal of the methods outlined in this paper is to choose experiments robust to distribution shift by leveraging what we know about existing sites. This comes with a trade-off: when our prior information is good, that is, highly prognostic, optimization does better. When our prior information is bad, randomization methods are more robust, and have better worst-case performance. I study this by simulation in Section 4.

Stratification compromises between random sampling and use of prior information and randomization (Thompson 2022). Stratification requires splitting the covariate space into strata, where we presume that the stratification occurs along dimensions of high treatment effect heterogeneity, so that the resulting strata capture meaningful variation in treatment response, and guarantee good coverage of the covariate space. It turns out that Optimal Transport methods can be interpreted as an optimal kind of stratification: the methods outlined below implement stratification where both partitions and representatives are simultaneously identified from the data. Essentially, we can think of the methods in this paper as a form of optimal, data-driven stratification. In practice, stratification often involves analyst-driven choices about what to stratify on. And with many covariates, it becomes less clear how the analyst should make high-dimensional stratification choices (though see (Tipton 2013a)).

A fundamental challenge in site selection is that we typically observe only a subset P of the universe of potential experimental sites $\mathscr{P}$. The distance between P and $\mathscr{P}$ is often unmeasurable, yet this is the population to which we ultimately wish to generalize. While no method can fully address this limitation, our approach provides robustness guarantees for deployment populations within a specified distance of the observed data.

In practice, researchers must play close attention to the informativeness of covariates collected in order to make decisions about site selection (Shpitser et al. 2012; VanderWeele et al. 2011; Stuart et al. 2013; Bicalho et al. 2022). Optimization can be a powerful tool to aid study design – if researchers engage in significant efforts to collect data at the planning stage. Randomization is still a preferable method if researchers do not have good information about sites.

1.2 Methodological Contributions

1.2.1 Optimal Transport and the Site Selection Problem

I use optimal transport theory to formulate the problem of selecting sites optimal for the Population Average Treatment Effect and Conditional Average Treatment Effect. Optimal transport is a rich body of applied mathematics with many possible applications in causal inference and machine learning (Villani 2003; Santambrogio 2015; Peyré et al. 2019). Optimal transport is concerned with the efficient shifting of mass between distributions, and gives rise to an intuitive notion of distance between distributions, the Wasserstein distance, which measures the shortest-cost transport distance between two distributions.

The Wasserstein distance quantifies how much "work" is required to transform one probability distribution into another, where work is measured as probability mass times the distance it must travel. Formally, for two distributions P and Q on a metric space, the p-Wasserstein distance is

$$W_p(P, Q) = \inf_{\pi} \left(\int \|x - y\|^p d\pi(x, y) \right)^{1/p}$$

where the infimum is taken over all transport plans π with marginals P and Q. Intuitively, if we think of P as describing the locations of piles of sand, Q as describing where we want to move that sand, and π as any given set of paths used to move sand from P to Q, the Wasserstein distance gives the minimum total cost of the move under the best routing from P to Q.

This metric is particularly well-suited for site selection because it directly captures the representativeness of selected sites. When we select experimental sites, we want them to "represent" the broader population in the sense that every population unit is adequately proxied by nearby selected sites. The Wasserstein distance formalizes this intuition: it measures how well a sparse set of selected sites can approximate a dense population by finding the optimal assignment of population units to selected sites while minimizing total "representation error."

For the special case of $p = 1$, this cost equals the population-weighted average distance that points must travel under the optimal assignment. For $p = 2$, we minimize the sum of squared distances, so W_2^2 equals the population-weighted average squared distance, and W_2 itself is the square root of this quantity—analogous to a population-weighted Euclidean distance. Just as standard deviation captures spread differently than mean absolute deviation, W_2 penalizes outliers more heavily than W_1: leaving any population point far from its nearest selected site contributes quadratically rather than linearly to the total cost.

In our site selection context, $W_1(P_X, S_X)$ measures the average distance from population units to their assigned experimental sites, while $W_2(P_X, S_X)$ is the root mean square Euclidean distance under optimal assignment, being more sensitive to ensuring no subpopulation is left too far from representation.

Many problems in causal inference can be formulated as optimal transport problems (Gunsilius 2025), and this is a rich vein of current and ongoing work (Torous et al. 2024).

I derive new upper bounds on the errors of the PATE and CATE estimator in terms of Wasserstein distances. By using the tools of optimal transport to analyze the Mean Squared Error of the PATE estimate, and the Precision in Estimated Heterogeneous Effect (Hill 2011; Shalit et al. 2017), I derive upper bounds for the PATE and CATE errors in terms of the Wasserstein distance (Theorem 14 and 15).

These bounds give us intuition about what our substantive goals are when choosing experimental sites for PATE and CATE estimation. When estimating the PATE, we seek a single number: the average treatment effect across the population. This means we want selected sites that, when averaged together, closely approximate the population average. Think of this as finding sites whose collective "center of gravity" matches the population's center.

When estimating the CATE, we want to estimate an entire function: how treatment effects vary across different covariate values. This requires accurate interpolation across the entire covariate space. We need sites spread throughout the population to avoid large gaps where we must extrapolate rather than interpolate.

These different goals lead to different selection strategies. For the PATE, the 1-Wasserstein distance naturally emerges because we care about average representation. Sites can compensate for each other, and modest coverage gaps in outlying areas are acceptable as long as the average is well-represented. For the CATE, the 2-Wasserstein distance emerges because outlying regions contribute quadratically to estimation error: leaving any subpopulation far from a selected site severely degrades our ability to estimate treatment effects in that region.

Both optimization problems seek to create *balanced partitions* of the covariate space that maximize representativeness, but they do so using different distance metrics that encode different notions of what "good representation" means.

For the PATE (1-Wasserstein), the optimization creates ℓ^1 Voronoi partitions—regions where each population point is assigned to its closest selected site in Manhattan distance. These diamond-shaped regions prioritize representing the population centroid and linear functionals, making this approach akin to balanced sampling on the class of all 1-Lipschitz functions (Deville et al. 2004). The ℓ^1 metric penalizes deviations linearly, leading to solutions that efficiently represent population means while ensuring no population unit is excessively far from its assigned representative.

For the CATE (2-Wasserstein), the optimization creates ℓ^2 Voronoi partitions—the familiar polygonal regions where assignment is based on Euclidean distance. The quadratic penalty of the ℓ^2 metric more heavily penalizes outliers, leading to solutions that provide more uniform coverage of the covariate space. This is better suited for function estimation tasks where we need to accurately interpolate treatment effects $\tau(x)$ across the entire support of X, rather than just estimating a single population parameter.

The geometric difference matters substantively: ℓ^1 partitions create diamond-shaped, axis-aligned regions, while ℓ^2 partitions create polygonal regions with perpendicular bisector boundaries. The ℓ^2 metric's quadratic penalty heavily penalizes outliers, leading to more equitable coverage across subgroups and ensuring no region of the covariate space lacks a selected site.

The ℓ^1 metric's linear penalty is more tolerant of coverage gaps in outlying areas, which is acceptable for PATE estimation where the goal is efficient estimation of a population average rather than function estimation across the entire support.

Why do metrics differ for PATE and CATE? The ℓ^2 penalty's outlier aversion leads to more equitable representation across subgroups. If you have minority subgroups located in outlying regions of the covariate space. This connects to fairness concerns in experimental design: the ℓ^2 penalty is more likely to include sites that represent minority or marginalized populations that might be located at the periphery of the covariate distribution.

Likewise, ℓ^2 outlier penalization is better for function estimation because CATE estimation requires accurate interpolation of $\tau(x)$ across the entire support of X. When you have gaps in coverage – that is, regions where the nearest selected site is very far awa – the function estimate $\tau(x)$ will be poor in those regions. This is particularly problematic with high levels of heterogeneity, because it makes it harder to reliably extrapolate $\tau(x)$ from distant points.

For the PATE, the resulting site selection strategy is akin to balanced sampling in survey methodology (Deville et al. 2004), where the goal is to balance on the infinite function class of *all* 1-Lipschitz functions, which includes linear functionals as a special case, but also smooth functions with bounded derivatives (see Appendix D).

For the CATE, the resulting sites selection strategy turns out to be approximately equivalent to optimal stratified sampling (see Appendix D), in which we *simultaneously* choose an optimal Voronoi partition of the covariate space, and representative sites from this partition.

In sum, the 1-Wasserstein distance (absolute distance) naturally arises for PATE estimation because we care about average representation, while the 2-Wasserstein distance (squared distance) is appropriate for CATE estimation because outlying regions contribute quadratically to function estimation error.

These upper bounds motivate a Mixed Integer Linear Program formulation of the PATE and CATE selection problems. Because our bounds contain Wasserstein distance terms, our objective then becomes to choose experimental sites that minimize the Wasserstein distance between the observed population of experimental sites and the selected sample of experimental sites, subject to a budget constraint of sites. Wasserstein distance minimization can be tractably reformulated in terms of Mixed Integer Linear Programs. These are straightforward to solve using commercial solvers like Gurobi. I develop software to implement this approach.

Empirical Performance versus Randomization and Optimization These optimization-based methods outperform simple random sampling when covariates are sufficiently informative about treatment effects, as I show via simulation in Section 4. A critical finding from our analysis is that optimization-based site selection requires observed covariates to explain more than approximately 50% of treatment effect variation ($R > 0.5$). This threshold has important practical implications: researchers should validate covariate informativeness before investing in optimization-based selection, as poorly informative covariates can lead to worse performance than randomization. These optimization methods can be understood as optimal stratified random sampling, and so the performance of optimal transport methods weakly upper bounds that of stratification methods (see 30).

1.2.2 Site Selection Under Distribution Shift

The observed population of sites may not represent the study population of interest. When planning experiments, researchers planning multi-site experiments face several possible sources of distribution shift. Experimental sites available for study may differ systematically from the target population due to selection bias (Allcott 2015); site characteristics may evolve between planning and implementation (Bansak et al. 2023); or covariates may be measured with error (Bound et al. 2001).

This is the problem of X-validity, in (Egami et al. 2023): when we are interested in generalizing from experimental samples in target populations: here, the problem is to engineer a sample that is X-valid with respect to many different populations.

Wasserstein Distributionally-Robust Optimization offers us tools to aid in decision-making under uncertainty. DRO methods address distributional uncertainty that arises when the deployment dataset differs from the training dataset (Duchi et al. 2020; Levy et al. 2020; Kuhn et al. 2024).

I extend the optimal transport framework using Wasserstein distributionally robust optimization (DRO). Rather than optimizing for the observed distribution, we can solve a more conservative problem that hedges against a range of plausible population distributions. Formally, our problem becomes:

$$\min_{S:|S|\leq K} \sup_{P' \in \mathcal{B}(P,\rho)} W_p(P, S)$$

where $\mathcal{B}(P, \rho) = \{P' : W_p(P, P') \leq \rho\}$ is an ambiguity set: the collection of all population distributions within radius ρ, measured in terms of the Wasserstein distance, around the empirical distribution. This provides worst-case performance guarantees when the true population lies within ρ of the observed data.

The ambiguity radius ρ is a scalar that represents the total "transportation budget" available to an adversary that seeks to perturb the observed distribution. Specifically, ρ bounds the total cost of moving probability mass in the covariate space, measured in the same units as the covariates themselves. For example, if covariates are standardized, then $\rho = 0.5$ allows Nature to move each population unit up to 0.5 standard deviations on average, or to make larger moves for some units while keeping others fixed, so long as the total transportation cost remains within budget. This provides worst-case performance guarantees when the true population lies within ρ of the observed data.

I solve the Wasserstein DRO site selection problem using a novel cutting-plane algorithm[2] that exploits the minimax game structure of the optimization problem. Formulating the DRO problem as a game theory problem directly suggests an algorithm for its implementation: the Researcher chooses a site selection; the adversary perturbs the observed data, subject to a budget on how far it can move points; the Researcher observes the adversary's new site selection and resolves the problem; and so on until neither the adversary nor the Researcher change their choices. (See Appendix D.2 for an explicit description of the equivalence.) Here, the Wasserstein DRO solution is interpretable as Nash Equilibrium in a game between Researcher and Nature; the algorithm proceeds by 'playing' the game between Nature and the Researcher until there are no further moves left. This removes the need to enumerate all elements of the (infinite) Wasserstein ball; instead, we identify only the set of adversarial best responses to a given site selection.

This game-theoretic cutting-plane approach is novel in the Wasserstein DRO literature. Existing methods for solving Wasserstein DRO problems typically rely on dual reformulations that convert the minimax problem into a single optimization (Esfahani et al. 2017), entropic regularization techniques that approximate the Wasserstein distance using Sinkhorn iterations to make the problem computationally tractable (Cuturi 2013), or moment-based approaches that replace Wasserstein constraints with simpler moment constraints (Gao et al. 2020).

[2] A cutting-plane algorithm solves optimization problems by iteratively adding constraints that eliminate infeasible regions. Rather than solving the full problem at once, the algorithm starts with a simplified version, finds a candidate solution, then checks if this solution satisfies all constraints of the original problem. If not, it adds a new constraint (a "cut") that rules out this solution and similar infeasible ones, then resolves the simplified problem. This process continues until the candidate solution satisfies all original constraints (Bradley et al. 1977).

The key insight of this approach is that we never need to characterize the full (infinite) ambiguity set $\mathcal{B}(P, \rho)$. Instead, we exploit the sequential structure: at each iteration, Nature reveals only the single adversarial distribution that is a best response to the current site selection, and we accumulate these best responses over iterations. This is computationally advantageous because we avoid both the approximation errors introduced by entropic regularization and the computational complexity of dual reformulations in high-dimensional covariate spaces.

I introduce a novel data-adaptive procedure for selecting the uncertainty radius in Wasserstein DRO problems. A separate technical contribution is the introduction of a novel data-driven calibration method for selecting the robustness parameter ρ. A fundamental challenge in applying distributionally robust optimization is choosing an appropriate robustness radius: too small provides insufficient protection against distribution shift, while too large yields overly conservative selections that sacrifice performance. Theoretical results provide guidance on how to select a robustness radius in the presence of sampling variability, based on the rate of convergence of empirical measures (Fournier et al. 2013; Blanchet et al. 2019b; Blanchet et al. 2021). However, it is difficult to formulate a theoretically principled way to choose a robustness radius in the face of unknown distribution shift beyond sampling variability: by design, we intend to guard against out-of-sample shifts, and so are limited in how we can use in-sample data to construct a plausible radius. This is because distribution shift in the wild induces *Knightian Uncertainty* (Knight 1921; Sunstein 2023): we cannot really know, without making assumptions, how much shift to guard against.

An alternative approach is to provide the option to guard against shifts that are benchmarked by the observed variation in the data. My procedure, detailed in Section 3.5, first constructs an empirical Wasserstein grid based on empirical distances in the covariate data. Intuitively, given any data set, there is a maximum radius beyond which an adversarial solution will not change. This motivates the heuristic procedure of 1) greedily searching for the maximum radius $\rho^{\max}$ and 2) performing adaptive grid search over the line $[0, \rho^{\max}]$. Site selection methods will produce different solution sets over this line: the goal is to identify when the output solutions exhibit small, moderate, and large differences from the baseline solution set. We can then define a series of ρ thresholds in terms of these different solution sets. Rather than requiring the user to specify ρ values, this procedure automatically generates ρ values that answer the question, "What would small, medium, and large distributional shocks look like for my specific dataset?." This makes DRO methods usable to practitioners without the need for arbitrary *priors* about what the appropriate radius of robustness should be.

Empirical performance I demonstrate the performance of these methods by reanalyzing Crépon et al (Crépon et al. 2015), who conduct a randomized microcredit experiment in Morocco, in which rural villages were randomized into receiving access to loans. I use as an outcome profits earned by individuals who did and did not take out the loan, and generate semi-synthetic treatment effects using observed covariates and a linear model. I first study the properties of site selections generated by my proposed methods, SPS, and random and stratified sampling on the full sample, evaluating the performance of these methods in terms of the MSE_{PATE} and the PEHE. I then implement a simulation study, in which treatment effects vary with signal strength (the informativeness of observed covariates), and in which I induce distribution shift by moving observed covariates away from their actual values. I show that my nonrobust methods outperform SPS under distribution shift, and in high-signal environments.

1.3 Summary of Proposed Methods

This paper introduces four methods for different practical use cases in site selection. First, the researcher should decide whether they are interested in PATE estimation or CATE estimation. Second, the researcher should decide how concerned they are about distribution shift: are they willing to pay 'the price of robustness' (Bertsimas et al. 2004) to trade-off accuracy in minimizing observed error against potential unobserved distribution shifts?

Four Site Selection Methods and Their Goals

Method	**Estimand**	**Objective**
$p = 1,\ \rho = 0$	PATE	Minimize MSE of Population Average Treatment Effect
$p = 1,\ \rho > 0$	PATE	Minimize worst-case MSE of PATE under distribution shift
$p = 2,\ \rho = 0$	CATE	Minimize PEHE (Precision in Estimation of Heterogeneous Effects)
$p = 2,\ \rho > 0$	CATE	Minimize worst-case PEHE under distribution shift

1.4 Related Literature

1.4.1 Multi-Site Experiments

In political science, the METAKETA initiatives coordinated by Evidence in Governance and Politics (EGAP) have systematically tested interventions across multiple countries, including voter information campaigns and electoral accountability (Dunning et al. 2019b; Dunning et al. 2019a), taxation and formalization policies, natural resource governance interventions (Slough et al. 2021), community policing programs (Blair et al. 2021; Blair et al. 2024), and women's action committees (Hyde et al. 2022). In economics, the Abdul Latif Jameel Poverty Action Lab (J-PAL) has pioneered large-scale coordinated evaluations, most notably the Graduation Program for the ultra-poor tested across six countries (Banerjee et al. 2015), and the Teaching at the Right Level initiative that has reached over 60 million students globally (Banerjee et al. 2017; Banerjee et al. 2016; Banerjee et al. 2007). Psychology has embraced multi-site replication through the Many Labs series, systematically testing the reproducibility of classic effects across dozens of laboratories (Klein et al. 2014; Klein et al. 2018; Ebersole et al. 2016; Ebersole et al. 2020). Public health has demonstrated the power of coordinated trials through initiatives such as the WHO SOLIDARITY trial for COVID-19 treatments involving over 14,000 patients across 35 countries (WHO Solidarity Trial Consortium et al. 2021; WHO Solidarity Trial Consortium 2022), the longitudinal Framingham Heart Study that established cardiovascular risk factors (Dawber et al. 1957; Kannel et al. 1961; Kannel et al. 1972), and the Women's Health Initiative examining hormone therapy effects across 161,000 participants (Rossouw et al. 2002; Anderson et al. 2004; Manson et al. 2013; Manson et al. 2024). Climate policy research has leveraged multi-jurisdictional implementation of carbon pricing mechanisms, particularly through analysis of the European Union Emissions

Trading System (Dechezleprêtre et al. 2023; Colmer et al. 2025; Klemetsen et al. 2020), to assess environmental interventions across political boundaries.

1.4.2 Site Selection in Causal Inference

(Egami et al. 2024) introduced explicit optimization methods for site selection in political methodology, and contributed significantly 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 et al. 2003; Alberto Abadie et al. 2010; Abadie et al. 2025). The 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. The default implementation contains a penalty term that additionally penalizes using outlying sites in the final selection.

The goal of this paper is to use a set of different technical tools to address the site selection problem motivated by (Egami et al. 2024). Whereas they use an approach based on synthetic controls intended to select experiments for the PATE, I i) show that the PATE and CATE have different optimization problems ii) use the theoretical resources of optimal transport to state and implement the minimization problem iii) use Wasserstein Distributionally-Robust Optimization to induce robustness to distribution shift.

(Tipton 2013a; Tipton 2013b) propose a cluster-then-stratify approach to site selection, which we study via simulation, and is weakly dominated by 2-transport, as I show in Appendix D

(Olea et al. 2024) solve the site selection problem, by defining it as the k-median problem. This is similar to the PATE transport solution, but the PATE solution implicitly imposes a balance constraint: that each site receive $\frac{1}{K}$ of the overall population mass. k-medians is not constrained in this way.

Optimal transport has a large number of possible applications for core causal inference tasks (Galichon 2016). Studying the changes-in-changes model (Athey et al. 2006), (Torous et al. 2024) use optimal transport methods to estimate control group trends over time, and apply this same transformation to predict what the treatment group would have looked like without intervention. In causal inference, (Bertsimas et al. 2023) applies distributionally robust optimization methods to the problem of learning treatment effects under unspecified confounding. They show that DRO can be interpreted as a form of sensitivity analysis. (Charpentier et al. 2023) propose using optimal transport methods to estimate counterfactual distributions, while (Dunipace 2022) use optimal transport methods to solve IPW-type problems (Hájek 1971; Horvitz et al. 1952; Ben-Michael et al. 2021).

1.4.3 Response Surface Methodology

The conceptual background of this paper is closely related to Response Surface Methodology, developed by (Box et al. 1951; Box et al. 1975; Box et al. 1987). In RSM, the goal is to choose experiments based on their location on the surface that determines how covariates map onto outcomes. This yields applied optimization problems, where we want to learn, say, the maximum of a given output function given inputs: this may correspond to an efficient configuration of industrial inputs, for instance. In our context, we can think of the treatment effect surface $\tau(X)$ as our response surface, and note that we want to choose experiments that are informative about the treatment effect surface, in a sense we will explore below.

1.5 Structure of Paper

Section 2 motivates the problem of site selection, and studies the case where the population of sites is observed, describes the assumptions needed to use covariates to select sites, states theoretical upper bounds on the downstream errors in estimating the PATE and CATE due to site selection, formulates the optimization problems associated with each estimand, and states algorithms to implement each procedure. Section 3 describes the application of Wasserstein DRO to the problem, motivates robust upper bounds, and describes a cutting-plane algorithm to implement Wasserstein DRO that leverages a game theoretic interpretation of the DRO problem. Section 4 studies the behavior of the site selection procedures by simulation. I study the performance of the methods against randomization as a function of signal strength, and show that these methods have good performance relative to randomization methods even for relatively weak signal strengths. I also characterize the robustness behavior of Wasserstein DRO empirically, and show that increasing the robustness radius in practice increases the coverage of the selected set. Section 5 reanalyses Crépon et al. (Crépon et al. 2015), an experiment in Morocco that randomized encouragement to access microcredit. I generate semi-synthetic treatment effects based on this data, and assess the behavior of the optimal transport and DRO methods compared to Synthetic Purposive Sampling and randomization methods as a function of problem size, signal strength, and distribution shift. Section 6 concludes.

2 Where to Experiment? The Problem of Site Selection

2.1 Overview of the Problem

Consider a researcher who is faced with a universe of sites $\mathscr{P}$, from which they must choose a subset S of sites, subject to the constraint that they can choose at most K sites.

The researcher's goal is to choose K sites that 'best represent' the population $\mathscr{P}$, in a sense that we will consider more specifically below.

We can formalize this by saying that the researcher must choose K sites that minimize a specific objective problem. The researcher is interested in the results of a downstream analysis of an experiment: they will eventually conduct an experiment and get an estimate of their population estimand of interest. The goal is to minimize the error of this estimate of the population quantity by selecting the 'best' sites at the planning stage of the experiment.

Remark 1. When $P = \mathscr{P}$, the researcher observes the full target population and faces a standard optimal transport problem: select sites S to minimize $W_p(P, S)$, the representation error. When $P \subset \mathscr{P}$, the researcher faces population uncertainty and must guard against the possibility that the observed sites P do not represent the true target population $\mathscr{P}$; the distributionally-robust method addresses this uncertainty.

1. The researcher defines a population of experimental sites $\mathscr{P}$, and chooses an estimand of interest (the $PATE$ or the $CATE$).

2. The researcher observes covariate information about a subpopulation of sites $P \subseteq \mathscr{P}$.

3. The researcher chooses a subset $S \subset P$ in which to run an experiment, where S contains at most K sites.

4. The researcher runs an experiment in the S sites, in-sample error is observed, and out-of-sample error is realized.

Figure 1: The Researcher's Site Selection Problem

2.2 Different objectives of Site Selection

The choice of objective function depends on the research context. First, the researcher must choose an estimand: they may be interested in the Population Average Treatment Effect (PATE), or the Conditional Average Treatment Effect (CATE).

Definition 2 (Population Average Treatment Effect (PATE))**.** $\mathbb{E}_{\mathscr{P}}[Y(1) - Y(0)]$

Definition 3 (Conditional Average Treatment Effect (CATE))**.** $\mathbb{E}_{\mathscr{P}}[Y(1) - Y(0)|X = x]$

For notational simplicity, I will write $\tau \equiv Y(1) - Y(0)$ and $\tau(x) \equiv Y(1) - Y(0)|X = x$, which are related by $\tau = \int \tau(x)dx$.

These represent fundamentally different statistical objectives that lead to different site selection strategies. In selecting the sites for the PATE, the downstream task is to estimate a *functional*: we seek to estimate a single number $\tau = E[Y(1) - Y(0)]$ that summarizes the average treatment effect across the population.

Estimating the CATE is a *function estimation* problem: we seek to estimate the entire function $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$ that describes how treatment effects vary across the covariate space.

This distinction has direct implications for site selection. For parameter estimation (PATE), we want sites that provide an efficient estimate of the population average. This requires representative sampling that balances coverage of different population subgroups. For function estimation (CATE), we want sites that enable accurate interpolation of $\tau(x)$ across the entire support of X. This requires broad coverage of the covariate space to minimize extrapolation error when predicting treatment effects at unobserved covariate values.

An advantage of increasing precision of the estimate of the CATE is that it should allow us to detect and characterize heterogeneous effects with greater accuracy. Reducing the estimation over the support of X allows us to do better inference on subgroups. For instance, with lower PEHE, we can more reliably test whether treatment effects differ between rural and urban sites, or identify which covariate combinations predict stronger treatment response.

While the CATE integrates to the PATE, minimizing estimation error for these two objectives requires different site selection strategies, as formalized in the upper bounds derived below.

In the exact/non-robust case, we are interested in finite sample PATEs and CATEs. But in the robust case, we are interested in the *extremal* PATE: the worst-case PATE given adversarial perturbations. Formally, if $P = \{x_1, \ldots, x_n\}$ are our observed sites, we consider the set of adversarially perturbed configurations $\mathcal{B}(P, \rho) = \{P' : W_p(P, P') \leq \rho\}$ and define the robust PATE as $\max_{P' \in \mathcal{B}(P,\rho)} \frac{1}{n} \sum_{i=1}^{n} \tau(x_i')$. This treats our observed sites as a single point in the configuration space of all possible site arrangements, with robustness defined over the ρ-neighborhood in Wasserstein distance[3]. Unlike superpopulation approaches that assume sites are sampled from some distribution, this framework provides local robustification around the specific observed configuration.

[3] We can think of this as an extension of the finite sample framework where populations are defined as orbits under bounded perturbations of observed data. This connects to the broader literature on minimax estimation and robust inference under bounded uncertainty, but applies the specific geometric structure of optimal transport to define perturbation sets. The configuration space perspective treats arrangements of experimental sites as points in an infinite-dimensional manifold equipped with Wasserstein distance, enabling local robustification without distributional assumptions (Amari 2012; Murray et al. 1993). I leave full development of this framework for causal inference to future work.

2.3 Site Selection When the Population is Observed

First, consider the case where the full population of sites is known to the researcher, the researcher has collected covariate information about all possible sites, and they can choose to run an experiment in any of those sites. This describes the case where $P = \mathscr{P}$. In this case, the expectations described in Definitions 2 and 3 are taken over the observed subpopulation P, because the population and subpopulation exactly coincide.

The below errors are 'downstream', because they are not realized until the analyst actually conducts the experiment. These quantities can be defined in advance of the experiment, and the infeasible problem that the analyst would like to solve can be stated.

2.3.1 Minimizing the Error of the PATE

For the PATE, we suppose that the researcher wants to minimize the Mean Squared Error of the downstream treatment effect estimate:

Definition 4. PATE problem when the population is observed

$$\min_S MSE_{\text{PATE}} = \min_S \mathbb{E}\left[\left(\frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}}\tau_i - \hat{\tau}^S\right)^2\right] \qquad \text{subject to } |S| \le K$$

Where the expectation is taken over randomness in treatment assignment and downstream estimation.

2.3.2 Minimizing the Error of the CATE

For the CATE, we suppose that the researcher wants to minimize the expected Precision in Estimation of Hetereogeneous Effect[4] (Hill 2011; Shalit et al. 2017).

Definition 5 (PEHE)**.**

$$PEHE = \int_X \left[\tau^P(x) - \hat{\tau}^S(x)\right]^2 dx$$

This gives us the researcher's minimization problem:

Definition 6 (CATE problem when population is observed)**.**

$$\min_S PEHE = \min_S \int_X \left[\tau^P(x) - \hat{\tau}^S(x)\right]^2 dx \qquad \text{subject to } |S| \le K$$

Because these errors are downstream, they are unobserved, and this exact minimization problem is infeasible. We can, however, use covariates to study feasible versions of these problems, and provide guarantees about how close the solution to these feasible problems are to the infeasible problems.

[4] Despite its name suggesting a measure to maximize, PEHE is an error metric where lower values indicate better performance. The name "Precision in Estimation of Heterogeneous Effects" is somewhat misleading as it measures estimation error rather than precision *per se*. Following the established literature (Hill 2011; Shalit et al. 2017), I retain the original terminology while noting that the goal is error minimization.

2.4 Assumptions Needed to Use Covariates To Select Sites

Assumption 7 (Observed Covariates Are Informative About Treatment Effects)**.**

$$\exists x, x' \in supp(X) \text{ such that } \tau(x) \neq \tau(x')$$

In words, treatment effects vary across the covariate space, making site selection based on covariates meaningful.

Assumption 8 (Common Mechanisms Across Sites)**.** *For sites $s \neq s'$:*

$$\mathbb{E}_{\mathscr{P}}[\tau(x, S = s)] = \mathbb{E}_{\mathscr{P}}[\tau(x, S = s')]$$

This stipulates that covariates have the same effect on treatment effect values across sites.

Assumption 9 (Lipschitz Continuity of τ)**.** *The treatment effect function $\tau : \mathbb{R}^d \to \mathbb{R}$ is Lipschitz continuous with constant L:*

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

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.

Assumption 10 (Independence of Experimental Design and Site Selection)**.** *Let Z_ℓ be the treatment assignment indicator and S_i be the site inclusion indicator. Then $Z_\ell \perp\!\!\!\perp S_i$.*

2.5 Optimal Transport: Some Tools and Definitions

In the next section, we use the tools of optimal transport to derive bounds on the errors of the MSE_{PATE} and $PEHE$. First, I introduce some terminology and notation, and a brief sketch of relevant concepts needed to state and solve our minimization problem. Optimal transport is a powerful methodological framework with broad application to problems in causal inference.

Optimal transport is concerned with moving mass between a source and a target in the most efficient way. An original motivating example, known as the Monge-Kantorovich Problem (Monge 1781; Ambrosio 2003; Vershik 2013), can be heuristically described as follows. Given a set of Parisian bakeries with specific production schedules and a set of cafes with specific consumption demands, located across Paris, what is the most efficient way to route bread from bakeries to cafes that minimizes the total transport distance? A transport map formalizes the idea of one possible solution to this problems: a collection of routes from bakeries to cafes, stored as a matrix. More formally, we have:

Definition 11 (Transport plan)**.** A **transport plan** between discrete distributions $P_X = \sum_{i=1}^n p_i \delta_{x_i}$ and $Q_Y = \sum_{i=1}^n r_i \delta_{y_i}$ is a matrix $\{\pi_{ij}\}_{(i=1,j=1)}^{(n,m)}$ such that $\sum_{i=1}^n \pi_{ij} = p_i$ and $\sum_{j=1}^m = r_i$.

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 ℓ^p distances as our cost function, so that $c(X, Y) = d_p(X, Y) = ||X - Y||^p$.

For $p = 1$, this gives us the absolute distance, and for $p = 2$, this is the squared distance between X and Y.

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

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

That is, if π^* minimizes the cost of transporting mass from P to Q measured in the p-norm,

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. The **p-Wasserstein distance** formalizes the notion of the shortest possible distance between P and Q, and is specified in terms of an optimal transport plan:

Definition 13 (p-Wasserstein Distance)**.** The p-Wasserstein distance between discrete distributions P and Q is given by:

$$W_p(P, Q) = \inf_{\pi} \sum_{i=1}^{n} \sum_{j=1}^{n'} \pi_{ij} ||x_i - y_j||^p$$

In our bakery example, this is defined in terms of the best possible solution to the routing problem between bakeries and cafes.[5]

I use the tools of optimal transport to derive upper bounds on the site selection problem: the Wasserstein distance is central to the theory that follows. I use P_X to denote the empirical distribution of covariates in the population, and S_X to denote the empirical distribution of covariates in the sample.

2.6 Upper-Bounding Errors Due to Site Selection

In order to minimize the error on the MSE_{PATE} and PEHE, we want to find a feasible upper bound on the problem that we can minimize via an optimization procedure. I derive two such bounds below. These bounds have the following properties:

The bounds do not depend on a specific model of treatment effects. That is, they are generically applicable to any site selection problem (as long as treatment effects vary smoothly with covariates).

The bounds make explicit the role of unmeasured heterogeneity. This allows us to be explicit about what our site selection tools can and cannot achieve, and to assess their performance under unmeasured heterogeneity empirically.

We can upper bound the errors of the MSE_{PATE} and the $PEHE$ by the 1-Wasserstein and 2-Wasserstein Distances between P_X and S_X, respectively.

[5] A political science versions of the optimal transport problem. Suppose we have a set of precincts and a finite set of campaign workers with different home locations. What is the most efficient way to assign campaign workers to precincts to minimize total distance traveled?

In each case we have a sensitivity parameter η_p, which measures how much the conditional distribution of unobserved covariates U differs between population P and selected sample S, given observed covariates X, which I call unmeasured heterogeneity.[6]

2.6.1 Upper-Bounding the MSE of the PATE

Theorem 14 (1-Wasserstein Bound on the MSE of the PATE)**.**

$$MSE_{PATE} \leq L^2 \cdot [W_1(P_X, S_X) + \eta_1]^2 + \sigma_S^2$$

Where $\eta_1 = \mathbb{E}_{P_X}[W_1(P_{U|X}, S_{U|X})]$ represents the degree of unmeasured heterogeneity, and σ_S^2 represents irreducible estimation error.

Note that η_1 conditions on observed covariates X, capturing only the residual unobserved variation. When unobservables are independent of observables ($U \perp\!\!\!\perp X$), we have $\eta_1 = W_1(P_U, S_U)$, the full distance between unconditional distributions. When unobservables are perfectly predictable from observables ($U = f(X)$), we have $\eta_1 = 0$. Thus η_1 automatically adjusts for observable-unobservable correlation, representing only the unobserved heterogeneity that remains after accounting for what we can measure.

2.6.2 Upper-Bounding the PEHE

Theorem 15 (2-Wasserstein Bound on the PEHE)**.**

$$PEHE \leq L^2 \cdot [W_2(P_X, S_X) + \eta_2]^2 + \sigma_S^2$$

Where $\eta_2 = \mathbb{E}_{P_X}[W_2(P_{U|X}, S_{U|X})]$ represents the effect of unmeasured heterogeneity, and σ_S^2 represents irreducible estimation error.

2.6.3 Discussion of Bounds

Why 1-Wasserstein for the PATE and 2-Wasserstein for the CATE? There is both a technical explanation and a substantive explanations.

In the proofs of Theorem 14 and Theorem 14, we get two upper bounds. In the first case, we note that the difference in estimated ATEs is a difference of linear functionals, and apply Kantorovich-Rubinstein to this difference. This is upper bounded by ther 1-Wasserstin distance.

In the second case, the PEHE is the integral of the squared pointwise errors in estimating $\tau(x)$ over X. The intuition is that squared pointwise errors $|\tau(x) - \hat{\tau}(x)|$ are bounded by $L||x - y||$ by Lipschitz continuity, so squared pointwise errors are bounded by $L^2||x - y||^2$; integrating both sides yields the 2-Wasserstein distance.

Another way to compare this is that in the PATE case, we are interested in linear function approximation, which yields linear penalties. In the CATE case, we are interested in an integral of squared pointwise errors – which has the same form as the 2-Wasserstein distance by construction.

[6] This captures two factors: *signal-to-noise ratio*, or how much treatment effects depend on unobserved U; and *unobserved covariate shift*: how the distribution of U between population and sample differs. The experimental planner observes only covariates and wants to know i) whether X is sufficient for treatment effects and ii) whether their sample is similar to the population on unobserved dimensions. This is unmeasured heterogeneity in the sense of site-level selection bias, rather than the more usual individual-level treatment assignment bias.

These bounds allow us to specify site selection as an optimization problem. The goal of these bounds is to find a feasible target for us to minimize via optimization. In both cases, our losses are upper-bounded by:

$$W_p(P_X, S_X) \text{ for } p \in \{1, 2\}$$

The p-Wasserstein distance between empirical distribution of covariates in the population and the sample. It is straightforward to minimize this quantity by choice of S using linear programming, as I show below.

Optimal site selections for the PATE and CATE differ. These bounds also help us to understand the difference in goals between selecting sites optimal for the PATE and selecting sites optimal for the CATE. The 1-Wasserstein distance places more weight on location, rather than variance; whereas the 2-Wasserstein distance more heavily penalizes outliers.

We have defined these bounds in terms of sensitivity parameters η_p, which allows us to study site selection under unobserved heterogeneity. Specifically, varying η_p through simulation, we can empirically assess when site selection methods outperform sampling – which, because they are randomized, are broadly robust to unobserved heterogeneity.

This also allows us heuristically to think about the role of data collection in the site selection process.

In the best case scenario, when we have perfect data collection, covariates are sufficient for treatment effects, so that $\eta_p = 0$, and site selection using observable covariates is a good idea.

In the worst case, observed covariates are completely uninformative about unobserved covariates, so that $U \perp\!\!\!\perp X$, and $\mathbb{E}[W_p(P_{U|X}, S_{U|X})] = \mathbb{E}[W_p(P_U, S_U)]$.

2.7 Similarity between optimization problems

In what follows, I will consider the more general problem of minimizing the p-Wasserstein distance, for $p \in \{1, 2\}$ on the understanding in that, when $p = 1$ we are minimizing an upper bound on the PATE, and when $p = 2$ we are minimizing an upper bound on the CATE. This considerably simplifies the exposition.

2.8 Minimizing The Upper Bounds Via Linear Programming

The bounds derived in the previous section give us clear objectives. If we want to select sites optimal for the PATE, we choose the sites S that minimizes the 1-Wasserstein distance between the empirical distribution of covariates in the selected sites S_X and the empirical distribution of the covariates in the population P_X. For the CATE, we select the sites that minimize the 2-Wasserstein distance.

We can formulate each site selection problem as a Mixed Integer Linear Program.

From Theorem 14 we now have the following optimization problem to minimize the upper bound on MSE_{PATE}:

$$\min_{S} \quad W_1(P_X, S_X) \quad \text{subject to} \quad |S| \leq K$$

To solve this problem, we can formulate it as a Mixed Integer Linear Program (MILP). Define the *site selection indicator* $s_i = \mathbb{I}\{s \in S\}$. Then, our optimization problem is:

MILP formulation for Site Selection Problem ($\rho = 0$)

$$\begin{aligned}
\min_{s,\pi} \quad & \sum_{j=1}^{|P|} \sum_{k=1}^{|P|} \pi_{jk} \|x_j - x_k\|^p \\
\text{subject to:} \quad & \\
& \sum_{j=1}^{|P|} s_j \le K && \text{(Site budget constraint)} \\
& \sum_{k=1}^{|P|} \pi_{jk} = \frac{1}{|P|} \quad \forall j \in P && \text{(Population marginal)} \\
& \sum_{j=1}^{|P|} \pi_{jk} = \frac{s_k}{\sum_{l=1}^{|P|} s_l} \quad \forall k \in P && \text{(Selected Subset's marginal)} \\
& \pi_{jk} \le s_k \quad \forall j, k \in P && \text{(Can only transport to selected sites)} \\
& \pi_{jk} \ge 0 \quad \forall j, k \in P && \text{(Non-trivial transport plan)} \\
& s_j \in \{0, 1\} \quad \forall j \in P && \text{(Site selection indicator is binary)}
\end{aligned}$$

Proposition 16. *For appropriate choice of p, minimizing the p-Wasserstein distance is equivalent to solving the above Mixed Integer Linear Program.*

Implementation details I solve the primal form of the problem directly. I implement this using the R Optimization Infrastructure (ROI) framework with multiple solver backends. The primary fallback solver is GLPK, which is freely available and provides reliable solutions for moderately-sized problems. For larger instances, the implementation calls Gurobi, a commercial solver that typically provides faster solution times and better numerical stability. The solver selection is automatic: the code attempts to use Gurobi if available, falling back to GLPK otherwise.

I use LP relaxation and warm starting to improve computational performance for larger problem instances. LP relaxation replaces the binary site selection variables $z_j \in \{0, 1\}$ with continuous variables $z_j \in [0, 1]$, converting the converting the MILP to a linear program that can be solved in polynomial time. This relaxed solution then provides a warm start for the exact MILP solver by initializing binary variables to rounded values of the relaxed solution. Runtime experiments show speedups of 2-40x depending on problem size. For problems with $n > 100$ sites, LP relaxation is used as the default since the relaxed solutions are often integral or near-integral for this problem structure.

2.9 Optimal Site Selections form Voronoi Partitions of the Covariate Space

The solutions to the PATE and CATE optimization problems induce Voronoi partitions of the covariate space.

Definition 17 (Voronoi Partition)**.** Given a finite set of sites $Y = \{y_1, \ldots, y_k\}$ in a metric space (X, d), the *Voronoi partition* of X with respect to Y is the collection of Voronoi cells $\{V_1, \ldots, V_k\}$ where

$$V_j = \{x \in X : d(x, y_j) \leq d(x, y_i) \text{ for all } i \neq j\}$$

Each cell V_j contains all points in X that are closest to site y_j under the metric d.

To understand these differences, we first derive the form of optimal site selections. Minimizing the p-Wasserstein distance $W_p(P, S)$ requires solving:

$$\min_{\pi} \sum_{i=1}^{|P|} \sum_{j \in S} \pi_{ij} \|x_i - x_j\|^p$$

subject to transport plan constraints. The optimal transport plan π^* assigns each population point x_i to the selected site that minimizes transportation cost:

$$\pi^*_{ij} = \begin{cases} \frac{1}{|P|} & \text{if } j \in \arg\min_{k \in S} \|x_i - x_k\|^p \\ 0 & \text{otherwise} \end{cases}$$

This creates a partition of the population where each selected site x_j receives all population points in its ℓ^p-Voronoi cell. By Definition 17, this is precisely a Voronoi partition of the population P with respect to the selected sites S under the ℓ^p metric:

$$V_j^{(p)} = \{x_i : \|x_i - x_j\|^p \leq \|x_i - x_k\|^p \text{ for all } k \in S\}$$

The choice of p determines the partition geometry. For the PATE, with $p = 1$, solution sets induce ℓ^1-Voronoi cells—diamond-shaped, axis-aligned regions where assignment is based on Manhattan distance. The linear penalty $\|x_i - x_j\|$ treats coordinate-wise deviations equally.

For the CATE, with $p = 2$, site selection is based on Euclidean distance. The quadratic penalty $\|x_i - x_j\|^2$ more heavily penalizes large deviations from any selected site.

The key difference lies in how these metrics handle population points far from any selected site. Under ℓ^2 optimization, the quadratic penalty makes outliers disproportionately expensive, forcing the algorithm to ensure no population region lacks nearby representation. Under ℓ^1 optimization, the linear penalty allows for moderate representation gaps if they improve overall efficiency.

PATE estimation (linear functional approximation) tolerates uneven coverage, while CATE estimation (pointwise function reconstruction) requires uniform spatial coverage to avoid interpolation gaps.

3 Site Selection Under Distribution Shift

In the previous section, we studied the problem of selecting sites optimal for the PATE and the CATE given observed information about the covariates. We can think of this as the full-information case: we assume that we have good knowledge of the data-generating process that determines treatment effects, and can have enough information to actually minimize the MSE of the PATE and the PEHE.

In practice, however, we might think that our data is imperfect, or measured with error. One way to formalize this notion is to say that our data is subject to distribution shift, or covariate shift.

Wasserstein Distributioanlly-Robust Optimization (DRO) is a set of methods for solving optimization problems with guarantees about the worst-case performance of a solution when the true underlying data distribution is a specified distance away from the observed data distribution (Esfahani et al. 2017; Blanchet et al. 2019a; Blanchet et al. 2024; Duchi et al. 2020; Kuhn et al. 2024).

To motivate Wasserstein DRO, we first motivate the notion of an ambiguity set:

Definition 18 (Ambiguity Set)**.** An ambiguity set of radius ρ around an empirical distribution P_n is the set of all distributions that are ρ-close to P in the p-Wasserstein metric.

$$B(P_n, \rho) = \{P \in \mathscr{P} \,:\, W_p(P_n, P) \leq \rho\}$$

Wasserstein DRO allows us to minimize the minmax risk over all candidate distributions in the ambiguity set.

3.1 Formalizing the Distributionally Robust Site Selection Problem

Conveniently, the formal results in the previous section specified upper bounds in terms of the Wasserstein distance from the empirical population to selected sample distributions.

The empirical idea here is analogous to that above: we minimize the Wasserstein distance between the sample and the population empirical distributions. Now, however, we explicitly take account of the fact that the empirical distribution P_X is not guaranteed to be a perfect representation of the underlying distribution that generated the data.

We can incorporate our uncertainty about the underlying distribution into our optimization problem via the ambiguity set. In particular, we want to minimize the worst-case risk[7], in the following formal sense:

Definition 19 (Distributionally-Robust Site Selection Problem)**.**

$$\min_{S\,:|S|\leq K} \sup_{P' \in B(P,\rho)} W_p(P', S)$$

Where, by plugging in $p \in \{1, 2\}$, we recover the site selection problems for the PATE and CATE respectively.

3.2 Wasserstein DRO as Game between Researcher and Nature

Distributionally-robust optimization has a conveneitn game-theoretic interpretation. Writing out the DRO problem again, we can see:

$$\underbrace{\min_{S\,:|S|\leq K} \overbrace{\sup_{P \in B(P,\rho)} W_p(P, S)}^{\text{Inner problem: Nature selects worst-case distribution}}}_{\text{Outer problem: Researcher selects sites}}$$

[7] Note that this differs from the sense of worst-case risk described in (Egami et al. 2024). They mean that they optimize an upper bound analogous to our results in the previous section; here I mean that we minimize the risk over an adversarially chosen distribution in the ambiguity set.

The inner sup is an action by adversarial Nature, to choose the worst-case distribution P, subject to the constraint that they can reallocate mass equal to at most ρ. In practice, this means that Nature can choose to relocate points adversarially (in practice, as outliers), selecting the worst-case distribution Q, and our result will still represent a valid upper bound on the chosen minimand. The outer minimization represents our best response to this adversarial perturbation. In short, ρ represents the budget of covariate shift that the researcher wishes to insure against.

3.3 Algorithm for Wasserstein DRO

This game-theoretic interpretation is not just a point of theoretical interest: it in fact motivates the algorithm I use to implement the DRO version of site selection.

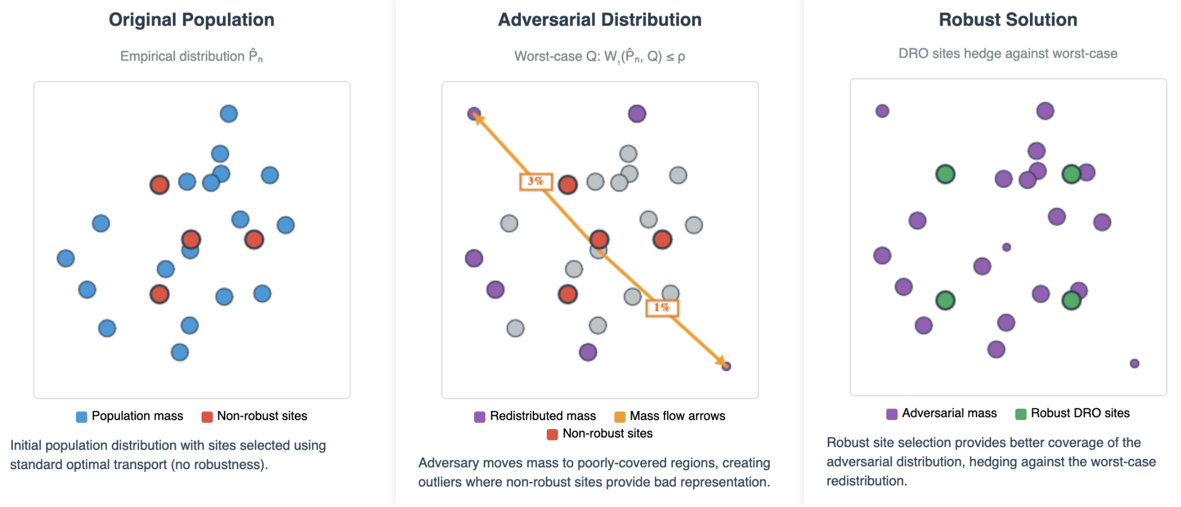
We have:

Algorithm 1 Heuristic Algorithm for Distributionally Robust Site Selection

```
Require: Site coordinates $X \in \mathbb{R}^{n \times d}$, number of sites $s$, robustness radius $\rho$, tolerance $\epsilon$
Ensure: Selected sites $S^*$, robust distance $W_p^*$
Initialize: Solve non-robust problem to get $S^{(0)}$
Set worst-case scenarios $\mathcal{Q}^{(0)} = \emptyset$, $t = 0$
while not converged do
    Given site selection $S^{(t)}$, Nature chooses an adversarial perturbation:
        $Q^{(t+1)} \in \arg\max_{Q:W_p(Q,P)\leq\rho} W_p(Q, S^{(t)})$
        Let $\text{UB}^{(t+1)} = W_p(Q^{(t+1)}, S^{(t)})$                ▷ Upper bound
    The adversarial perturbation is stored in memory:
        $\mathcal{Q}^{(t+1)} \leftarrow \mathcal{Q}^{(t)} \cup \{Q^{(t+1)}\}$
    Researcher minimizes site selection error against all observed adversarial perturbations:
        $S^{(t+1)} \in \arg\min_{S:|S|=s} \max_{Q\in\mathcal{Q}^{(t+1)}} W_p(Q, S)$
        Let $\text{LB}^{(t+1)} = \max_{Q\in\mathcal{Q}^{(t+1)}} W_p(Q, S^{(t+1)})$      ▷ Lower bound
    if $\text{UB}^{(t+1)} - \text{LB}^{(t+1)} < \epsilon$ then
        break                                ▷ Gap is small: solution is near-optimal
    end if
    $t \leftarrow t + 1$
end while
$S^* \leftarrow S^{(t+1)}$
return Selected sites $S^*$ and robust distance $W_p^*$
```

The ambiguity set $B(P, \rho)$ is built constructively out of Nature's best responses to the Researcher's site selections. We do not need to enumerate all elements of the Wasserstein ball, which is an infinite set; we need only enumerate the adversarial perturbations that increase the Researcher's observed loss.

Proposition 20. *The solution S^* of Algorithm 1 is ϵ-close to the minimizer of the Wasserstein DRO site selection problem.*



Original Population

Empirical distribution $\hat{P}_n$

Initial population distribution with sites selected using standard optimal transport (no robustness).

Adversarial Distribution

Worst-case Q: $W_1(\hat{P}_n, Q) \leq \rho$

Adversary moves mass to poorly-covered regions, creating outliers where non-robust sites provide bad representation.

Robust Solution

DRO sites hedge against worst-case

Robust site selection provides better coverage of the adversarial distribution, hedging against the worst-case redistribution.

Figure 2: How Wasserstein DRO applied to Site Selection works in practice. Given an initial site selection, the adversary perturbs the probability mass assigned to observed sites. The researcher chooses a new site selection, and the adversary responds. This process continues until the selection is stable.

3.4 Intuition: What kind of robustness is Distributional Robustness?

Incorporating the robustness parameter ρ allows to describe new bounds on our estimates. This gives us the *robust* upper-bounds:

$$\sup_{Q \in B(P_n,\rho)} MSE_{\text{PATE}}(Q,S) \leq L^2 \cdot (W_1(P,S) + \rho + \eta_1)^2 + \sigma^2$$
$$\sup_{Q \in B(P,\rho)} PEHE(Q,S) \leq L^2 \cdot (W_2(P,S) + \rho + \eta_2)^2 + \sigma^2$$

Where these guarantees are given over a Wasserstein ball[8] around the observed distribution. DRO ensures that the solution is robust to distribution shift – that is, robust to changes in the distribution of observed covariates. We can also think of this as measurement error: our solution should be robust to a specified degree of mismeasurement ρ. This is in contrast to the parameter η_p, which represents outcome model error due to unmeasured heterogeneity. procedure.

The Wasserstein ambiguity set $B(P_X, \rho) = \{Q : W_p(Q, P_X) \leq \rho\}$ contains *all* distributions that can be reached by moving the observed covariate distribution's mass by at most ρ units. Each distribution Q in this set represents a different way our observed site characteristics could be wrong: measurement error, temporal drift, or systematic misrepresentation of the target population.

The core idea is that an adversary creates gaps in how representative our sample is by strategically relocating probability mass. ρ represents uncertainty about where the population is located in covariate space. The larger the budget ρ, the more mass the adversary can relocate to regions poorly served by our specific site selection.

The "worst case" distribution Q^* is the one that maximally exploits differences in site characteristics. For example: the worst-case distribution might concentrate all mass in rural extremes, given an initial selection of urban sites.

By optimizing against the worst case, we obtain a site selection that is robust to *every* distribution in the ambiguity set. This is because our selection must perform well against the adversary's best response: which means it performs at least as well against any other distribution the adversary could have chosen. In this sense, Wasserstein DRO provides insurance against *all possible covariate shifts of magnitude* ρ, representing all the ways we could have mismeasured the true site characteristics.

The robustness parameter ρ controls shifts in observed covariates, while our bounds include an additive term $\eta_p = E_{P_X}[W_p(P_{U|X}, S_{U|X})]$ capturing unobserved heterogeneity. This formulation already accounts for observable-unobservable correlation: when X and U are independent, η_p equals the full Wasserstein distance between unconditional distributions; when they are perfectly correlated, η_p approaches zero. The additive structure $(W_p(P,S) + \rho + \eta_p)$ thus correctly separates robustness to observable shifts (ρ) from residual unobserved heterogeneity after conditioning on observables (η_p).

In practice, the adversary shifts the entire observed distribution, including components correlated with unobservables. This means that choosing ρ based on empirical variation provides implicit protection against correlated unobserved factors. This is not explicitly stated in the bounds, which are conservative and describe the pessimistic case where X

[8] Constraining shifts to be within a Wasserstein ball simply limits the total mass that can be moved around, and specifies a cost – either an ℓ^1 or ℓ^2 penalty, depending on the estimand – for doing so.

and U are orthogonal. When X and U *are* correlated, the effective robustness exceeds what the additive bound suggests, as the ρ-ball constrains both observable variation and its correlated unobservable components.

3.5 Procedure for Choosing Robustness Parameter ρ

How should one choose the degree of robustness in practice? Experiments detailed in 4.2.2 show that, at high levels of ρ, the algorithm can increase the coverage of the solution set.

Choosing a robustness radius is a practical problem in Wasserstein DRO, because it is not clear i) what the robustness radius means in real terms and therefore ii) how practitioners should think about radius selection.

I propose an automated, data-driven method, that benchmarks levels of distribution shift against variation observed in the data, and gives users a choice of levels of shift to guard against.

The idea is to do grid search over values of ρ, and evaluate the stability of the solution set as ρ changes.

Define the Jaccard similarity:

Definition 21. Jaccard similarity $J(S_1, S_2) = \dfrac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$.

The core idea is to calculate the Jaccard similarity between the initial, non-robust, baseline solution – the site selection chosen by the exact optimal transport methods – and the solutions chosen given different adversarial budgets $\{\rho_1, \rho_2, \dots, \rho_{\max}\}$.

The Jaccard similarity $J(S^{(0)}, S^{(\rho)})$ compares the non-robust baseline solution $S^{(0)}$ (obtained with $\rho = 0$) to increasingly robust solutions $S^{(\rho)}$. This measures how much the optimal site selection *changes* as we demand more robustness

This Jaccard radius selection procedure chooses robustness parameters in Wasserstein DRO by constructing an empirical Wasserstein grid from pairwise distances between all sites in the covariate space.

The algorithm performs a greedy search to identify $\rho_{\max}$, the maximum radius beyond which adversarial solutions cease to change meaningfully. Starting from the non-robust baseline solution $S^{(0)}$, the procedure solves the DRO problem at empirical distance quantiles and tracks solution stability using the Jaccard similarity.

When the Jaccard similarity falls below 0.5, indicating that half the sites in the robust solution differ from those in the non-robust solution, the algorithm terminates the search and sets $\rho_{\max}$. A refined grid search over $[0, \rho_{\max}]$ then maps the solution path, allowing automatic classification into four robustness levels: none ($\rho = 0$), moderate ($75 - 90\%$ solution overlap), high ($50 - 75\%$ overlap), and maximum ($< 50\%$ overlap). This procedure generates ρ values that answer the question: "What would small, medium, and large distribution shifts look like for my specific dataset, given observed variation on observable covariates?"

Algorithm 2 Data-Adaptive Robustness Radius Selection via Jaccard Similarity

```
Require: Site coordinates $X \in \mathbb{R}^{n \times d}$, number of sites $s$, Wasserstein norm $p$, grid resolution $n_{\text{grid}}$
Ensure: Robustness levels $\{\rho_{\text{moderate}}, \rho_{\text{high}}, \rho_{\text{maximum}}\}$
Compute empirical distance matrix: $D_{ij} = W_p(\delta_{x_i}, \delta_{x_j})$ for all $i, j \in [n]$
Extract pairwise distances: $\mathcal{D} = \{D_{ij} : i \neq j\}$
Solve baseline problem: $S^{(0)} \in \arg\min_{S:|S|=s} W_p(\hat{P}_n, S_X)$        ▷ Non-robust case
Initialize: $\rho = 0$, $\mathcal{J} = \emptyset$, converged = False        ▷ Greedy search for $\rho_{\max}$
for $\rho \in$ quantiles($\mathcal{D}$, $[0.1, 0.2, \ldots, 0.9]$) do        ▷ Empirical grid
    Solve DRO problem: $S^{(\rho)} \in \arg\min_{S:|S|=s} \sup_{Q:W_p(Q,\hat{P}_n)\leq\rho} W_p(Q, S_X)$
    Compute Jaccard similarity: $J^{(\rho)} = \frac{|S^{(0)} \cap S^{(\rho)}|}{|S^{(0)} \cup S^{(\rho)}|}$
    Store: $\mathcal{J} \leftarrow \mathcal{J} \cup \{(\rho, J^{(\rho)})\}$
    if $J^{(\rho)} < 0.5$ or plateau detected then        ▷ Solutions diverge significantly
        $\rho_{\max} \leftarrow \rho$, break
    end if
end for        ▷ Grid search
Define grid: $\mathcal{G} = \{\rho_1, \rho_2, \ldots, \rho_{n_{\text{grid}}}\}$ over $[0, \rho_{\max}]$
for $\rho_k \in \mathcal{G}$ do
    Solve DRO problem: $S^{(k)} \in \arg\min_{S:|S|=s} \sup_{Q:W_p(Q,\hat{P}_n)\leq\rho_k} W_p(Q, S_X)$
    Compute Jaccard similarity: $J^{(k)} = \frac{|S^{(0)} \cap S^{(k)}|}{|S^{(0)} \cup S^{(k)}|}$
end for
$\rho_{\text{moderate}} \leftarrow \min\{\rho_k : J^{(k)} \in [0.75, 0.90]\}$        ▷ Small perturbation
$\rho_{\text{high}} \leftarrow \min\{\rho_k : J^{(k)} \in [0.50, 0.75]\}$        ▷ Moderate perturbation
$\rho_{\text{maximum}} \leftarrow \min\{\rho_k : J^{(k)} < 0.50\}$        ▷ Large perturbation
return $\{\rho_{\text{moderate}}, \rho_{\text{high}}, \rho_{\text{maximum}}\}$
```

The intuition behind the procedure is that there must be a maximum adversarial perturbation budget ρ^{max}, such that, for any $\rho > \rho^{max}$ the 'most robust' site selection does not change. This is because variation in the data is finite. This motivates the following heuristic procedure: quickly find ρ^max, and then do adaptive grid search on the interval $[0, \rho^{max}]$, where we may sequentially add refinements in order to ensure that we collect enough site solutions $S(\rho)$ to be able to estimate $J(S(\rho), S(\rho'))$ for a large number of pairs.

Once we have this similarity measure for enough points, we can compare the observed similarities $\{J(S(\rho_i), S_(\rho_j))\}_{ij}$ and rank them, giving us a set of solution sets with decreasing similarity. We then output a set of three increasing ρ values such that the solutions at each ρ have decreasing similarity to the baseline solution $\rho = 0$. This ensures that we have solution sets that increase in dissimilarity to the nonrobust solution as the radius ρ increases.

4 Simulations: Randomization versus Optimization, and Solution Sets

4.1 Overview of Simulations

The simulations address three questions: (1) How do solution sets differ between 1-Wasserstein (PATE)and 2-Wasserstein (CATE) optimization? (2) How do site selection solutions change as the robustness parameter ρ increases? and (3) How do optimization methods compare to selection methods based on random sampling and stratified sampling?

First, I provide visual characterizations of solution sets generated by different objectives (Section 4.1), illustrating how the 1-Wasserstein objective (PATE) trades off between central location and coverage while the 2-Wasserstein objective (CATE) more heavily penalizes leaving any region uncovered. These illustrations use synthetic covariate data to build intuition about the geometric properties of each method and how the robustness parameter ρ affects solution set coverage.

Second, I conduct systematic simulation studies to evaluate when optimization-based selection outperforms randomization approaches (Section 4.2). For these performance comparisons, I generate candidate populations with covariates $X_s \sim \mathcal{N}(0, I_5)$ and site-level treatment effects $\tau_s = \sqrt{1-\eta^2} f(X_s) + \eta\varepsilon_s$, where $\eta \in [0,1]$ controls the signal-to-noise ratio: the fraction of treatment effect variation unexplained by observed covariates. I compare simple random sampling (uniform selection), stratified random sampling (k-means clustering followed by within-stratum sampling), and the optimization methods across varying signal strengths. The key finding is that optimization methods dominate when $\eta < 0.7$ (equivalently, when observable covariates explain more than 50% of treatment effect variation).

4.2 Characterizing site selection solution sets

4.2.1 PATE vs CATE selection patterns

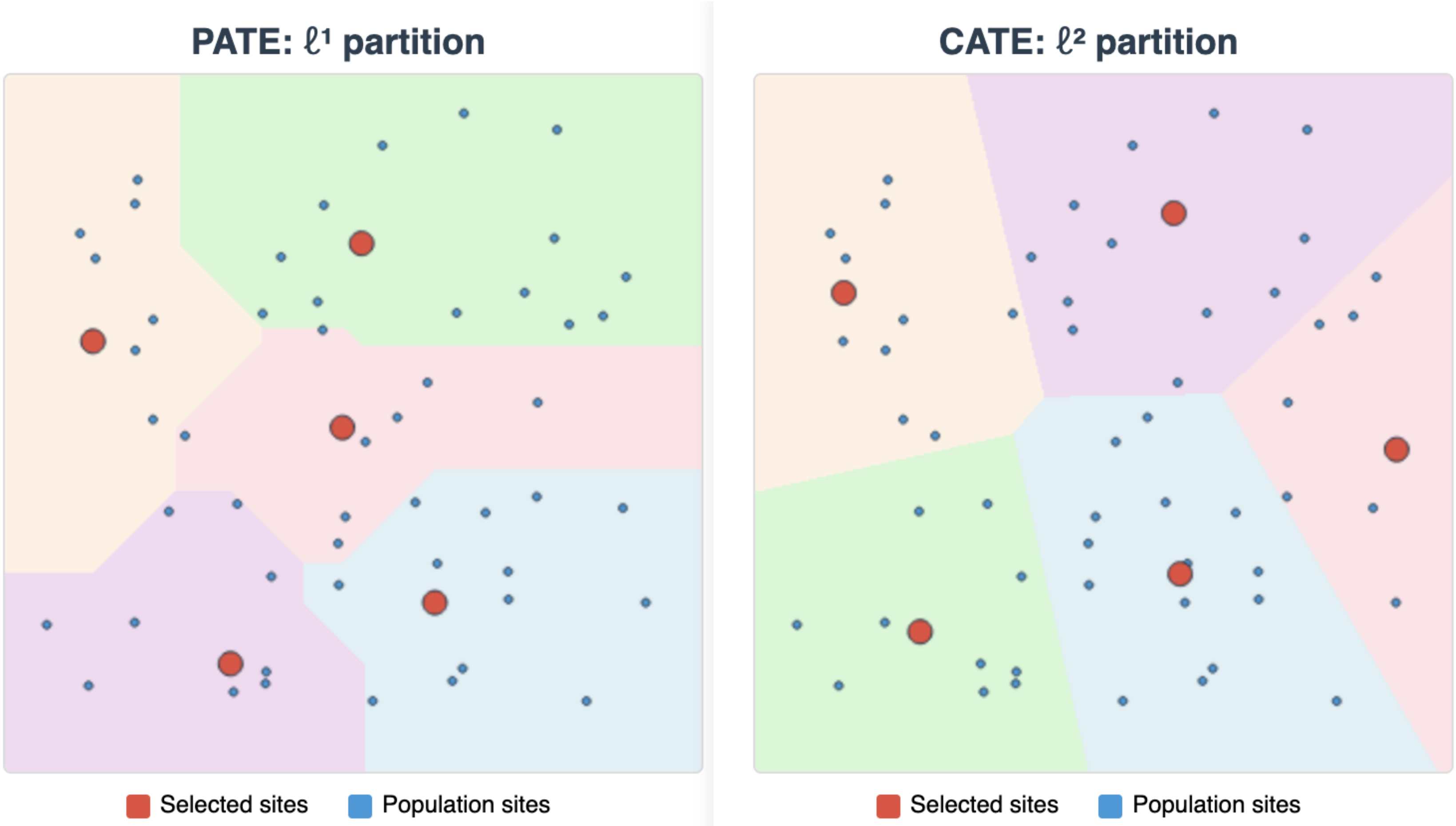


Figure 3: **Illustrative solution sets for the PATE and the CATE.** Both methods simultaneously solve for optimal Voronoi partitions of the covariate space, and optimal representatives within that partition. They differ with respect to the p-norm used to solve the problem. We can understand both methods as optimal versions of stratified sampling (see Appendix D); the ℓ^1 norm places more weight on location, while the ℓ^2 norm places more weight on minimizing the variance of the site selection. is therefore an optimal version of stratified sampling. The difference is between choosing points that well represent the support of a *function*, $\tau(X)$, which requires good coverage of the space of X, versus choosing points that well represent a *functional* $\mathbb{E}[\tau(X)]$, which requires choosing sites that provide good coverage of a single point, the population centroid.

The above bounds show that there are different site selection objectives for the PATE and the CATE. In the PATE case, we care about the 1-Wasserstein Distance, and in the CATE case the 2-Wasserstein distance.

Recall that the 1-Wasserstein distance contains the absolute norm, and the 2-Wasserstein distance is a function of the ℓ^2 norm. This entails that while the cost of increasing distance is linear in the 1-Wasserstein case, the cost of increasing distance from unselected points to selected points is quadratic in the difference of distances.

This should penalize selections that are far away from unselected points more in the 2-Wasserstein case, leading to a more compact set for the 1-Wasserstein solution and a larger set for the 2-Wasserstein solution.

This is intuitively appealing in the causal inference context, since the 1-Wasserstein distance is associated with the PATE, where our best guess of the PATE is the centroid of our observed sites. The CATE problem involves estimating a function over the support of X, and so, intuitively, we would want a solution set with improved coverage over the support of X.

To test these theoretical predictions, I generate synthetic datasets with known covariate distributions and compare the geometric properties of optimal site selections under both objectives. The simulation uses $|P| = 30$ candidate sites distributed across a two-dimensional covariate space, from which $K = 5$ sites are selected.

In practice, for small-sized problem instances, the solution sets are fairly similar. This is because, for sufficiently well-behaved data, site selections that minimize the 1-Wasserstein distance also minimize the 2-Wasserstein distance and vice versa. This behavior is analogous to that of Least Absolute Deviations versus Ordinary Least Squares – while using the ℓ^1 distance rather than the ℓ^2 distance does in fact produce different solutions, these solutions may not be qualitatively different.

However, as the dimensionality and complexity of the covariate space increases, the differences become more pronounced. The CATE solutions exhibit systematically larger convex hull areas and greater dispersion, consistent with the goal of function estimation over the support of the space rather than centroid approximation.

In our causal inference context, the practical implication is that, for small sized problem sets, solution sets that are optimal for the PATE are likely also to be optimal for the CATE. The CATE objective, in principle, prioritizes coverage over the space, so that we can learn $\mathbb{E}[\tau|X = x]$ for a large support X. The PATE objective prioritizes coverage of the center, so that we learn the average location with high probability. In practice, however, good coverage of the space implies good coverage of the average, and a solution that minimizes absolute distance from selected sites to non-selected sites will also provide good coverage of the support of the covariates.

4.2.2 Effect of robustness parameter ρ on solution set coverage

The robustness parameter ρ controls the budget allocated to the adversary in the distributional robustness problem. As ρ increases, the DRO framework hedges against increasingly severe distribution shifts by selecting more dispersed site configurations. This section demonstrates how robustness considerations systematically alter the geometry of optimal selections.

To illustrate this behavior, I solve the DRO problem across a range of ρ values and track the evolution of site selection patterns. The simulation uses a two-dimensional covariate space with $|P| = 30$ candidate sites, selecting $S = 5$ sites at different robustness levels.

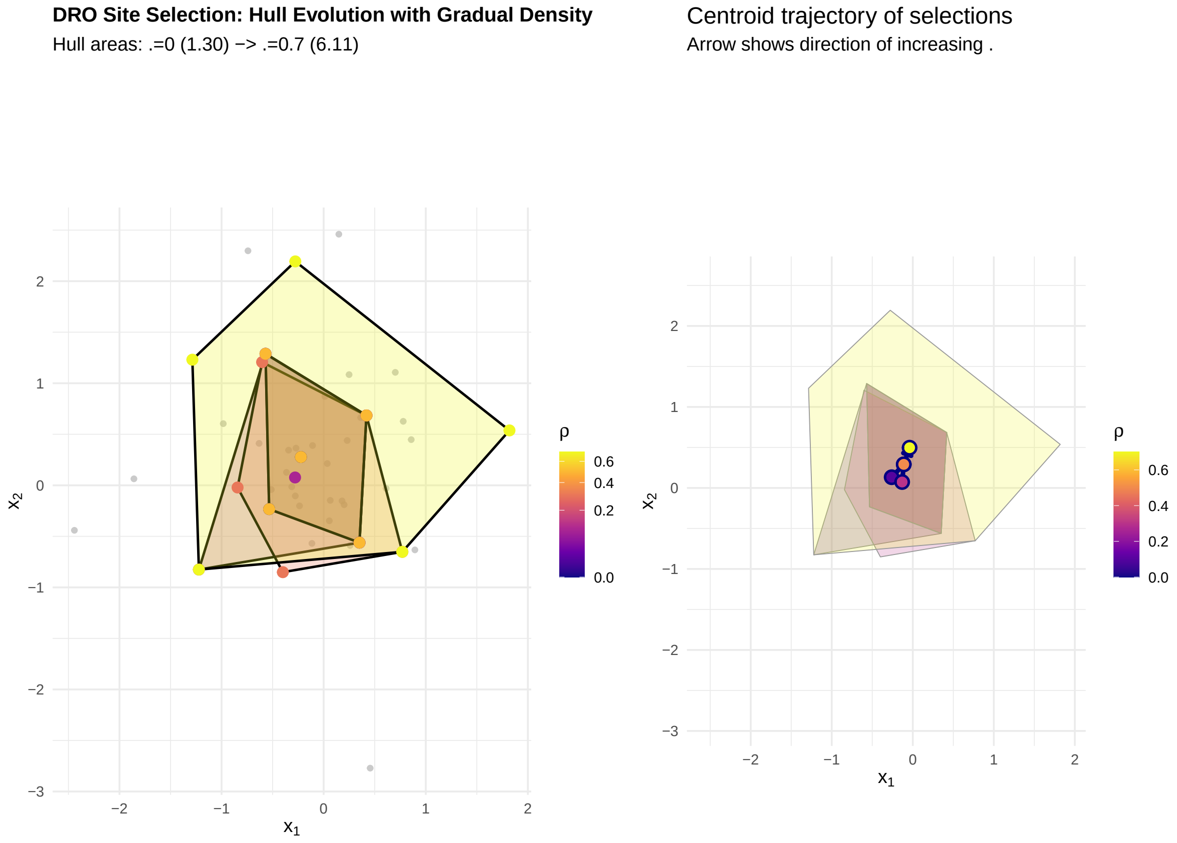


Figure 4: **As ρ increases, site selections become less compact.** The convex hull area expands from 1.30 to 6.11 as ρ increases from 0 to 0.7, demonstrating the systematic trade-off between optimality and robustness. The centroid trajectory shows how the selection focus shifts (marginally) away from the population center toward broader coverage as distributional uncertainty increases. This is a visualization of the 'price' of robustness' – some amount of drift in our point estimate of the PATE.

This robustness-coverage trade-off has important implications for experimental design under uncertainty. Researchers facing potential distribution shift should choose ρ values that balance the benefits of robustness against the costs of suboptimal site allocation. The Jaccard radius selection procedure, described in Section 3.5, provides an automated way to select this radius, with implications for the size of the hull selected.

4.3 Comparing random sampling and optimization methods via simulation

4.3.1 PATE: Optimization vs randomization

Randomization is minimax optimal for experimental selection when the researcher has no prior information about experiments (Kallus 2020). We are essentially using prior information, in the form of covariates, to choose sites, and would expect that the quality of our site selection improves as covariates become more informative.

The key question is: at what threshold of covariate informativeness do optimization methods cease to provide benefits over simpler approaches? This threshold determines the practical applicability of the optimization procedures.

To evaluate this, I run a simulation in which the site selections are evaluated over a grid of η values, where η controls the degree of unmeasured confounding, as in the upper bounds derived above. There is a mild reparameterization, as η is now defined on the support $[0, 1]$, with the interpretation that $\eta = 0$ implies that covariates are sufficient, and there are no unobserved determinants of treatment effect, while $\eta = 1$ implies that covariates are completely uninformative about treatment effects, and the optimization methods are essentially fitting to noise.

The simulation generates treatment effects using the parameterization detailed in Appendix B.1, which allows systematic variation of signal strength while maintaining realistic correlation structures between covariates and outcomes.

The goal is to compare the optimization procedures to 1) complete randomization, in which sites are selected at random and 2) stratification, in which k-means is first used to separate the sites into strata, and sites are then sampled from the k clusters. This is the procedure suggested in (Tipton 2013a).

These represent two different assumptions about our prior information. Complete randomization implies that we have no information about potential outcomes from covariates. Stratification implies that we have some information about covariates: we know that some covariates are important enough that we should condition our randomization on them. Stratification can be understood as a compromise between complete randomization and optimization approaches: it is a constrained randomization approach.

The simulation study confirms our theoretical expectations: optimization methods perform better when covariates are informative up to $\eta \approx .7$. We can translate $\eta \approx .7 \implies R^2 \approx .5$. The Crépon study below has an R^2 of .66, which would mean we had good enough covariates to consider optimization-based selection methods.

This breakdown point has important practical implications. Researchers should validate covariate informativeness before relying heavily on optimization-based site selection.

This suggests a straightforward moral: optimization methods outperform random assignment when covariates are sufficiently informative about potential outcomes.

4.3.2 CATE: Optimal stratified sampling

I show this result formally in Appendix D. The intuition is that to select sites that provide optimal coverage of the support of the function, 2-Wasserstein transport *simultaneously* selects an optimal partition and optimal representatives of the space. This is in distinction to stratification, where optimal representatives are identified given a partition. Hence, 2-Wasserstein transport provides a weak lower bound on the error of the stratified sampling solution.

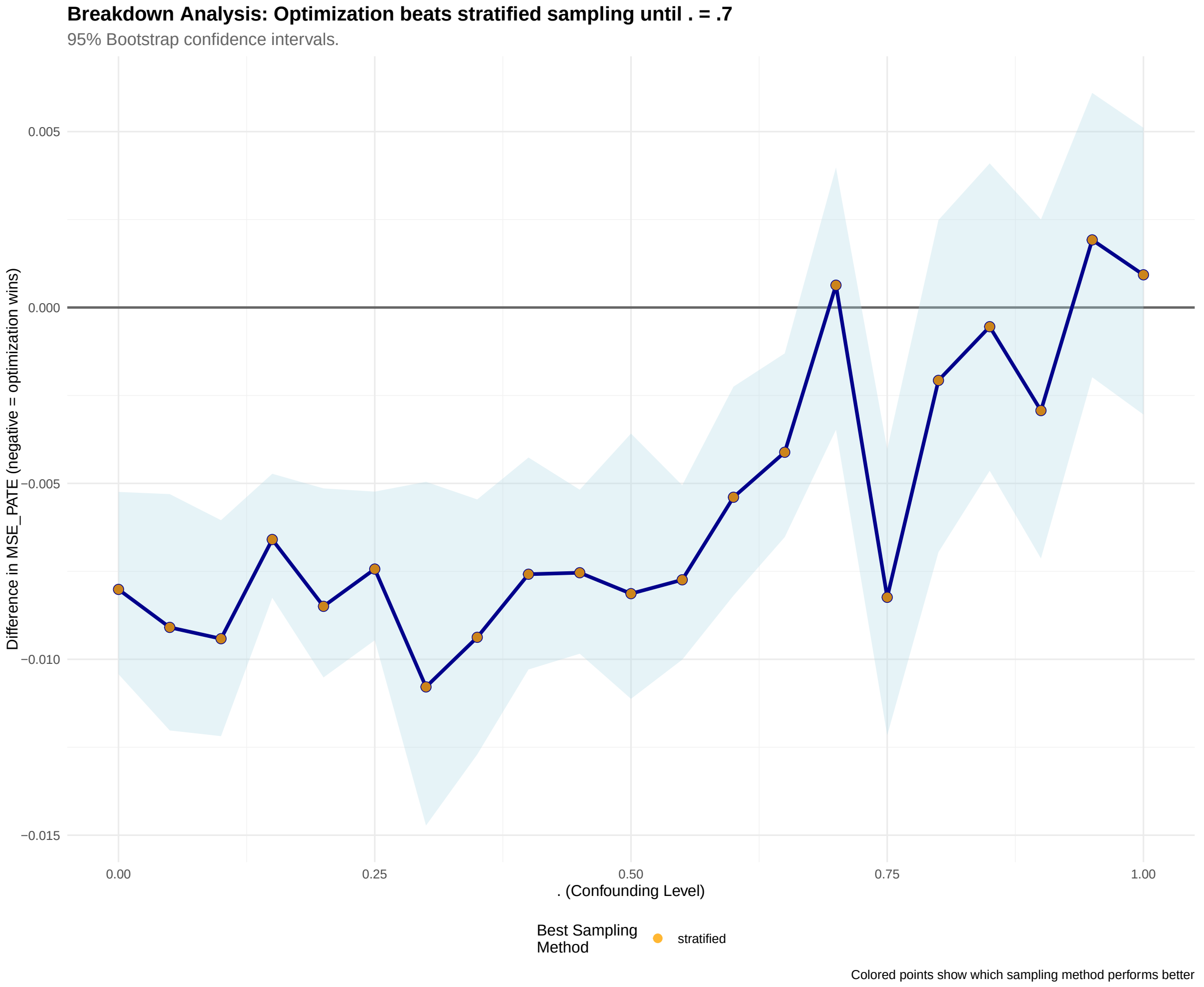


Figure 5: Performance of PATE optimization method as unmeasured heterogeneity increases. The optimization advantage diminishes as η approaches 0.7, beyond which randomization weakly dominates. Error bars represent 95% confidence intervals based on 1000 simulation replications.

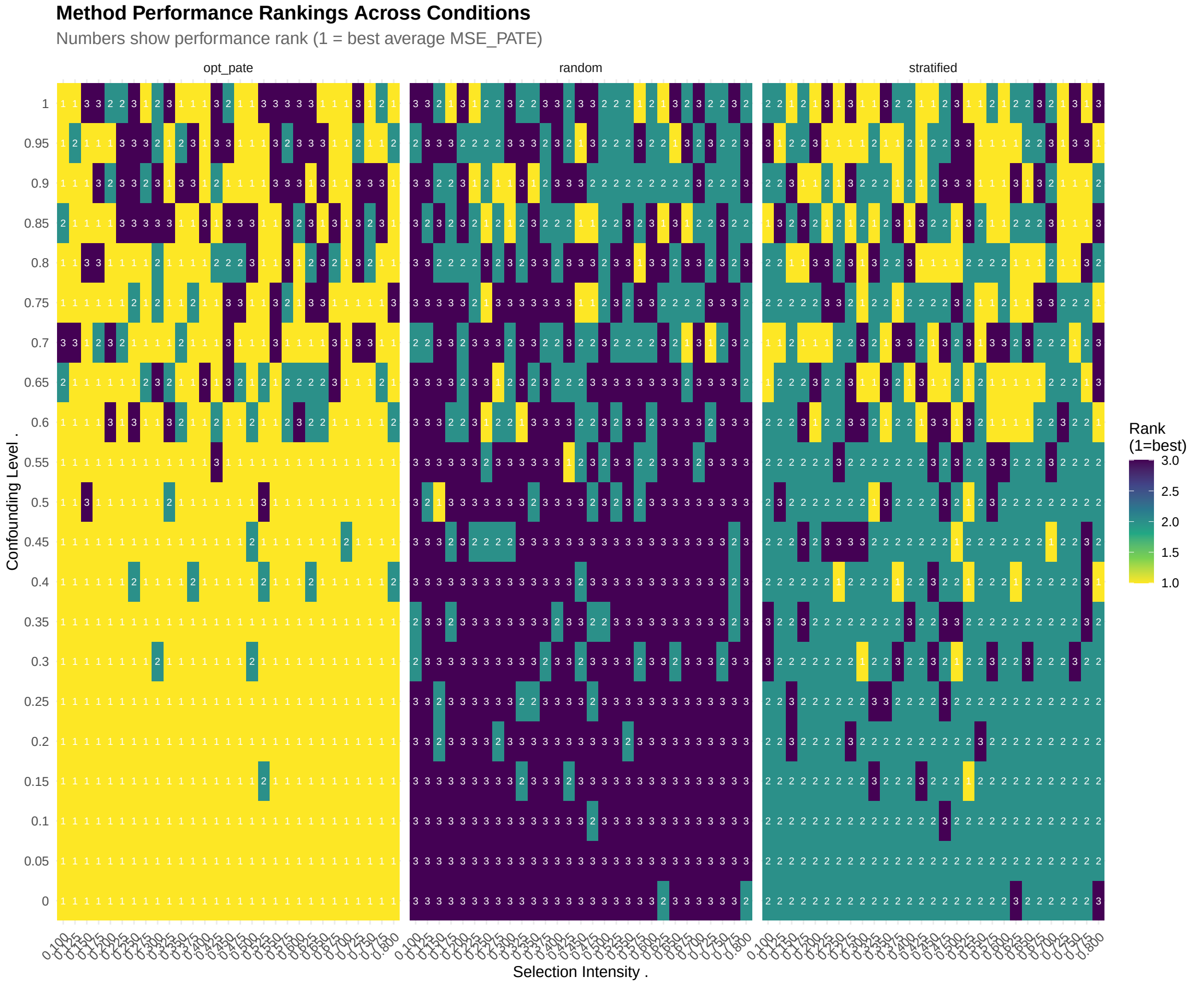


Figure 6: **Optimization breaks down versus random sampling when between** $50-90\%$ **of treatment effect variance comes from unobserved factors** (95% **CI**). η parameterizes the degree of unmeasured heterogeneity. In figure 4a) we can see that optimization outperforms stratification until $\eta > .7$. 95% bootstrapped confidence interval for this breakdown point is $[.7, .95]$. In figure 4b),optimization dominates when signal strength is high (η is close to 0); with stratification beating randomization otherwise.

This equivalence provides both theoretical insight and computational advantages. Theoretically, it shows that the optimal transport framework provides a lower bound on the error of stratificaton. Computationally, it allows us to leverage established stratified sampling algorithms as benchmarks.

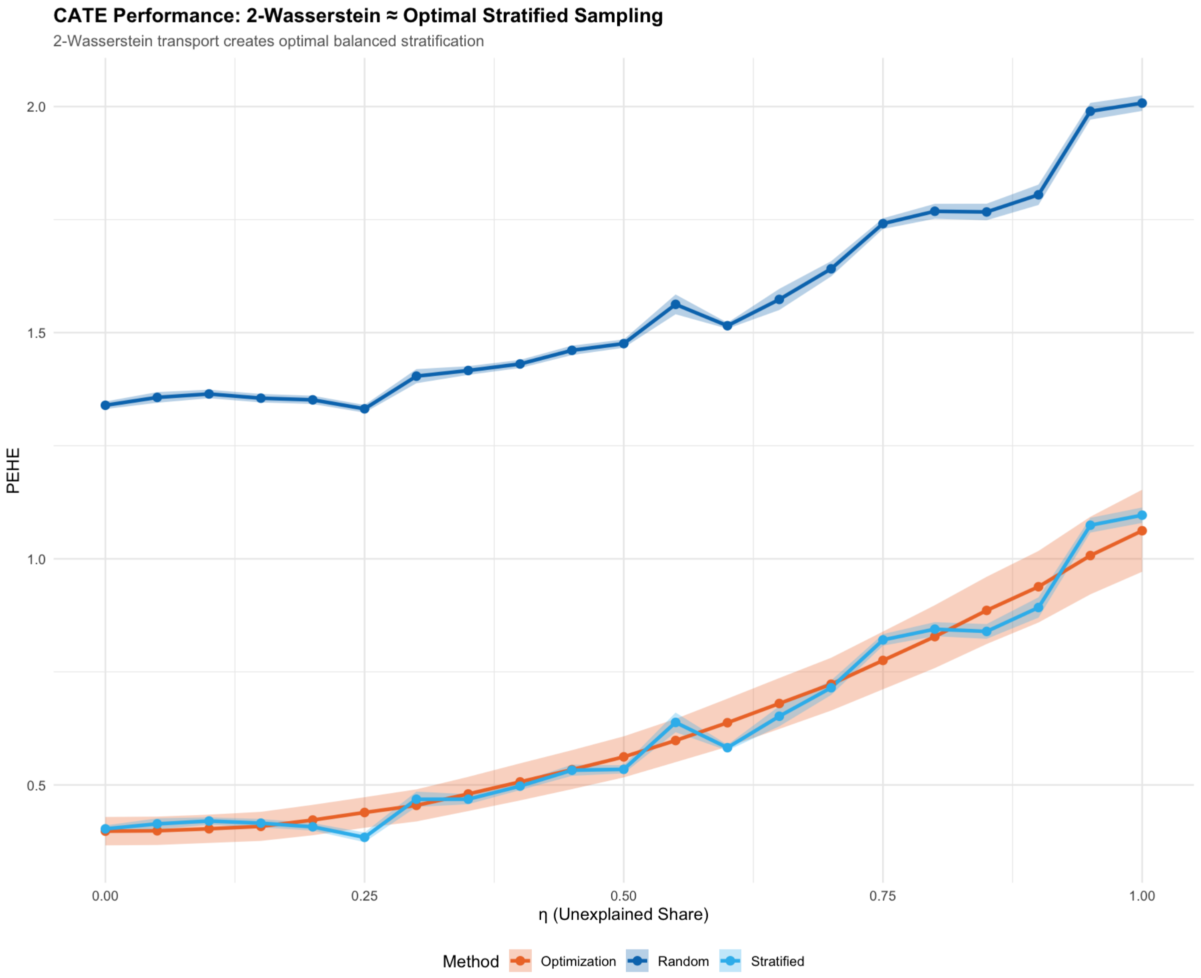


Figure 7: The CATE optimization method performs roughly equivalently to optimal stratified sampling. Both methods achieve similar PEHE values across different signal strength levels, confirming the theoretical equivalence.

The connection to stratified sampling arises through the geometric structure of optimal transport solutions. As shown in Section D, the 2-Wasserstein optimal transport problem induces a Voronoi partition of the covariate space, where each selected site serves as the representative for all population points in its Voronoi cell. Formally, for optimal sites $\{s_1^*, \ldots, s_K^*\}$, the induced partition is $\mathcal{V}_j = \{x \in \mathcal{X} : ||x - s_j^*|| \leq ||x - s_k^*|| \text{ for all } k \neq j\}$.

This Voronoi structure emerges endogenously from the transport optimization—it is not imposed as a constraint but rather characterizes the optimal assignment of population points to their nearest selected sites. The 2-Wasserstein objective ensures these cells are balanced while minimizing within-cell variance. Thus, while the method is fundamentally different from classical stratification, it produces a partition with similar geometric properties: spatially contiguous regions with designated representatives that minimize representation error.

The equivalence between 2-Wasserstein transport and stratified sampling is conceptual, not procedural. Stratified sampling sequentially (1) partitions the covariate space via k-means, then (2) randomly samples within partitions. Optimal transport chooses representatives that minimize distance from each selected site to their closest representation; the partition structure is implicitly defined from the solution, rather than being the result of constraints.

4.4 When should you use optimization methods in practice?

Our simulations demonstrate that optimization methods outperform random selection when observable covariates explain at least 50% of treatment effect variation ($R^2 > 0.5$ or $\eta < 0.7$). This threshold is achieved in many policy-relevant settings—for instance, the Crépon et al. microcredit study analyzed below has $R^2 = 0.66$. Researchers should use optimization when they have strong priors or evidence about treatment effect predictors; otherwise, stratified random sampling provides a robust alternative that partially leverages covariate information while maintaining unbiasedness. This also highlights the important of collecting prognostic covariates prior to experimental deployment (Bicalho et al. 2022).

How can researchers assess whether their covariates are sufficiently informative (i.e., $\eta < 0.7$ or $R^2 > 0.5$) without running the full experiment? Several approaches are available: (1) *Prior experiments*: Use treatment effect estimates from similar interventions to assess how much variation is explained by observable characteristics. For instance, in educational interventions, prior multi-site trials can reveal whether school-level characteristics predict treatment effects. (2) *Pre-treatment outcomes*: When available, the relationship between covariates and pre-treatment outcomes provides a lower bound on their relationship with treatment effects. (3) *Domain expertise and theory*: In many fields, accumulated knowledge suggests which characteristics drive heterogeneity—for example, baseline health status in medical trials or institutional capacity in policy interventions. (4) *Pilot studies*: Small-scale pilots across diverse sites can estimate treatment effect heterogeneity before committing to the full experiment. When such evidence suggests $R^2 < 0.5$, stratified random sampling provides a robust compromise between optimization and randomization.

5 Simulation: Crépon et al. (2015).

Crépon et al. (2015) studied the effects of a randomized microcredit intervention in Morocco. They considered a population of 162 villages, which were randomized into 81 matched pairs. Treatment consisted of an encouragement campaign to take out credit from Al Amana banks: "door-to-door campaigns, meetings with current and potential clients, contact with village associations, cooperatives, and women's centers, etc." (129).

These villages that were randomized into treatment were a population of sites that were on the periphery of catchment areas of existing branches: the goal was to assess whether taking up microcredit had an impact on a number of economic variables.

In this simulation, we take household self-employment activity profits as the outcome. We estimate the effect of treatment, site-level and individual covariates on profits, and estimate synthetic treatment effects for every individual in the sample using observed information. Sites are selected on the basis of aggregate-level site data, and we then estimate the error in terms of MSE_{PATE} and $PEHE$ for each site selection. A more detailed description of the simulation procedure can be found in Appendix B.

5.1 Simulation Procedure

Our simulation consists of four main components: baseline parameter estimation, synthetic data generation, site selection method application, and performance evaluation.

5.1.1 Baseline Parameter Estimation

We begin by estimating the predictive power of village-level covariates for treatment effects using the empirical Cr'epon data. We aggregate individual-level data to site level and estimate site-specific treatment effects. We then regress these site-level treatment effects on baseline village characteristics to estimate the signal-to-noise ratio, finding that the empirical signal strength $R^2 = .66$.

5.1.2 Synthetic Data Generation

For each simulation run, we:

1. Sample village-level covariates from the empirical distribution
2. Apply the trained treatment effect model to predict site-level effects
3. Add controlled noise to achieve target signal-to-noise ratios
4. Generate individual-level outcomes consistent with site-level parameters

The noise level is calibrated such that the proportion of treatment effect variance explained by covariates matches the specified signal strength (0.3, 0.66, or 0.9).

5.1.3 Distribution Shift Implementation

We implement distribution shift by modifying the covariate distributions of candidate sites relative to the deployment population. Shift magnitude is expressed as multiples of the empirical Wasserstein distance observed in the original Crépon data. For shift magnitude $\varsigma \in \{.4, .6, .9, 1.7, 3.4\}$, we transform candidate site covariates such that:

$$W_2(P_X, P_{\text{shift}}) = \varsigma \times W_2$$

This approach grounds simulation conditions in realistic population variation. The simulation is run for two signal-to-noise ratio levels: .3, .9. These correspond to a low signal and high signal case respectively.

The actual degree of treatment effect variance explained by observed covariates in the Crepon data is .66: we therefore benchmark our simulation conditions against the actual predictiveness of covariates observed in the data. We have three cases: the low-signal case, the benchmark case, and the high-signal case. This is helpful, because it is useful to consider the behavior of these methods in the context of a realistic social science study, with naturalistic data collection.

We also benchmark distribution shift against observed variation in the data. We calculate the actual variation in the data, and study the behavior of the methods. Because this is a simulation study, however, we can induce plausible degrees of distribution shift that are also benchmarked against naturalistic observed shifts in the data. This is done by estimating shifts based on the empirical Wasserstein distances in the data; and inducing distribution shift as a percentage of these observed shifts.

5.1.4 Method Implementation

We implement five site selection methods:

- **Random**: Uniform random selection from candidate sites
- **SPS**: Synthetic Purposive Sampling using convex hull optimization
- **Optimal Transport (Non-Robust)**: Wasserstein distance minimization without robustness
- **Wasserstein DRO**: Distributionally robust optimization with uncertainty radius ρ
- **Stratification**: K-means clustering followed by within-cluster random sampling

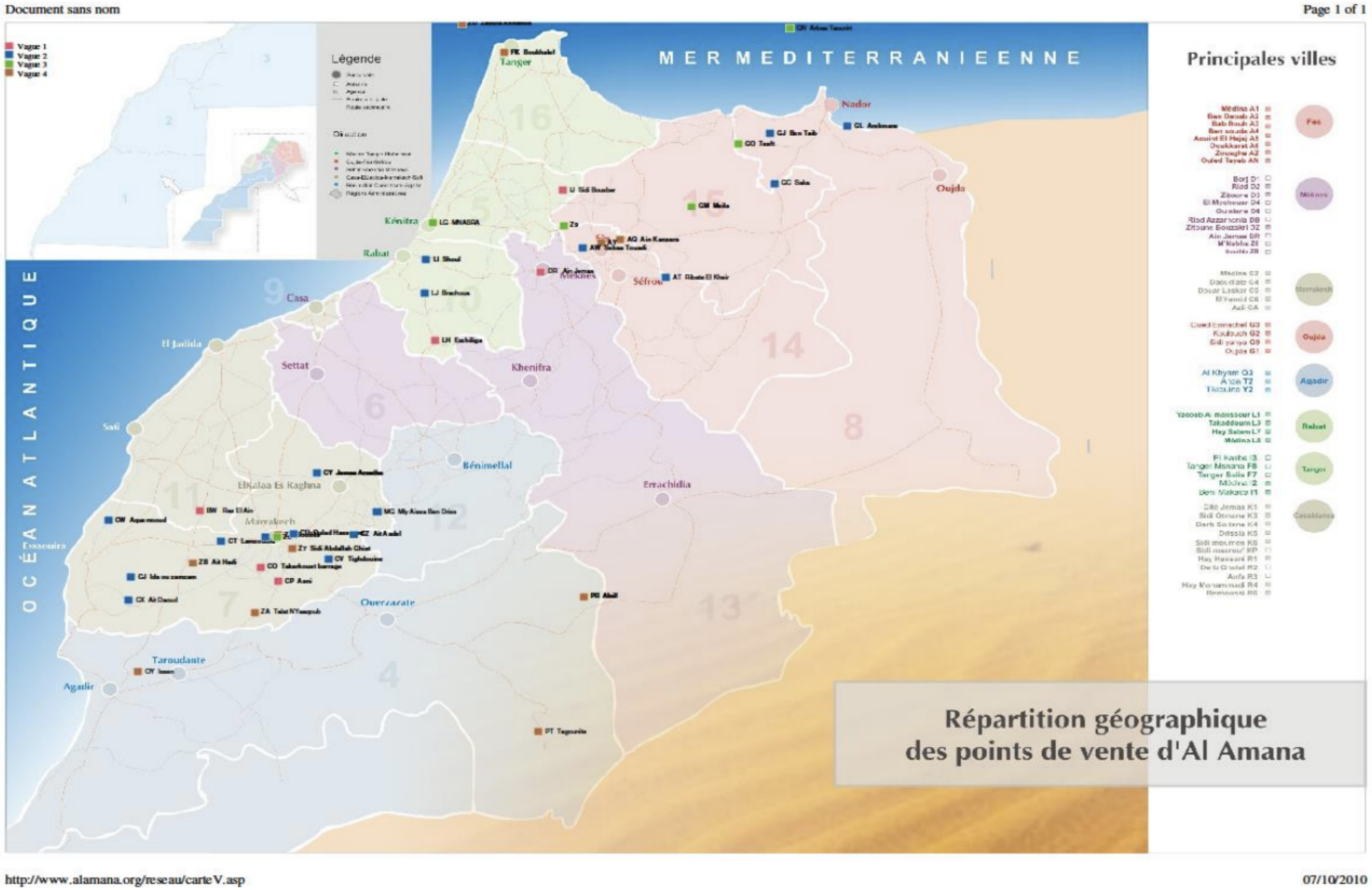
Each method selects K sites from a pool of N candidate sites, with $(N, K) \in (20, 4), (25, 5)$ corresponding to realistic experimental scales.

5.1.5 Performance Evaluation

For each site selection, we estimate PATE and CATE using standard methods and compare to ground truth values calculated from the complete synthetic population. Performance metrics include:

- $MSE_{\text{PATE}} = (\hat{\tau}\text{PATE} - \tau\text{PATE}^{\text{true}})^2$
- $PEHE = \mathbb{E}[(\hat{\tau}(X_i) - \tau^{\text{true}}(X_i))^2]$

We conduct 500 simulation runs per scenario to ensure stable performance estimates.



Document sans nom

Page 1 of 1

http://www.alamana.org/reseau/carteV.asp

07/10/2010

Figure 8: Sites selected in Crépon et al.

Problem Size	Signal	Shift	Winner	Advantage
20 choose 4	0.3	0.0	SPS	71.9%
20 choose 4	0.3	0.4	SPS	48.0%
20 choose 4	0.3	0.6	SPS	9.0%
20 choose 4	0.3	0.9	SPS	45.4%
20 choose 4	0.3	1.7	Wasserstein DRO	39.8%
20 choose 4	0.3	3.4	Wasserstein DRO	49.9%
20 choose 4	0.9	0.0	Optimal Transport	43.6%
20 choose 4	0.9	0.4	Optimal Transport	21.8%
20 choose 4	0.9	0.6	Optimal Transport	30.1%
20 choose 4	0.9	0.9	Optimal Transport	34.2%
20 choose 4	0.9	1.7	Wasserstein DRO	10.1%
20 choose 4	0.9	3.4	Wasserstein DRO	10.9%
25 choose 5	0.3	0.0	Optimal Transport	3.4%
25 choose 5	0.3	0.5	Wasserstein DRO	3.5%
25 choose 5	0.3	1.0	Wasserstein DRO	4.7%
25 choose 5	0.3	1.3	Wasserstein DRO	7.9%
25 choose 5	0.3	1.6	Wasserstein DRO	12.0%

Table 2: **Results: Error in estimation of the $MSE_{\mathbf{PATE}}$** by method result. Best-performing method over all simulation runs is reported here. Advantage is % reduction in error of the MSE_{PATE}.

5.2 Results

The simulation results demonstrate three main patterns. First, site selection method choice produces larger performance differences for PATE estimation than for CATE estimation. Second, the relative performance of methods depends on signal strength and problem size. Third, distributionally robust methods become preferred under realistic degrees of distribution shift.

5.2.1 PATE Performance Results

For PATE estimation, performance advantages range from 3.4% to 71.9% . Under low signal strength (0.3), SPS dominates when distribution shift is minimal, but Wasserstein DRO becomes optimal when shift exceeds 1.7 times empirical variation. Under high signal strength (0.9), Optimal Transport methods generally outperform alternatives, except under large distribution shift where DRO maintains advantages.

5.2.2 CATE Performance Results

For CATE estimation, performance differences between methods are substantially smaller, with most advantages below 1%. This pattern holds across signal strength and shift conditions, indicating that CATE performance depends more on fundamental signal-to-noise constraints than on site selection method choice. Our results show that site selection for the PATE is qualitatively different to site selection for the CATE. In Appendix D,

Table 3: PEHE Performance Summary Table

Problem Size	**Signal**	**Shift**	**Winner**	**Advantage**
20 choose 4	0.3	0.0	Optimal Transport	0.9%
20 choose 4	0.3	0.4	Optimal Transport	0.3%
20 choose 4	0.3	0.6	Tie	$< 0.1\%$
20 choose 4	0.3	0.9	Tie	$< 0.1\%$
20 choose 4	0.3	1.7	Wasserstein DRO	0.7%
20 choose 4	0.3	3.4	Wasserstein DRO	0.7%
20 choose 4	0.9	0.0	Tie	$< 0.1\%$
20 choose 4	0.9	0.4	Tie	$< 0.1\%$
20 choose 4	0.9	0.6	Optimal Transport	0.3%
20 choose 4	0.9	0.9	Tie	$< 0.1\%$
20 choose 4	0.9	1.7	Tie	$< 0.1\%$
20 choose 4	0.9	3.4	Wasserstein DRO	0.5%
25 choose 5	0.3	0.0	Tie	$< 0.1\%$
25 choose 5	0.3	0.5	Optimal Transport	0.1%
25 choose 5	0.3	1.0	Wasserstein DRO	0.1%
25 choose 5	0.3	1.3	Optimal Transport	0.1%
25 choose 5	0.3	1.6	Tie	$< 0.1\%$

Table 4: **Error in estimation of the $PEHE$ by method result.** Best-performing method over all simulation runs is reported here, differences of less than .1% reported as a tie.

I show that there are theoretical equivalences between optimal transport methods and familiar survey sampling approaches.

5.2.3 Optimal Transport methods perform better for medium-to-large site selection problems

SPS methods have an advantage in the $\binom{20}{4}$ case under low signal strength, but are dominated by Optimal Transport methods for the larger problem size of $\binom{20}{5}$.

The transition point likely occurs because convex hull approaches suffer from dimensionality limitations while optimal transport methods handle larger optimization spaces efficiently.

5.2.4 Optimal Transport methods perform better in high-signal strength conditions

Optimal transport methods strictly dominated in the signal = .9 case. This was true for both the original and shifted problems, with performance advantages over SPS ranging from 10.1% to 43.6%.

5.2.5 DRO methods perform better for larger distribution shift levels

The crossover point where DRO methods become preferred occurs at shift levels of 1.7 times observed empirical variation. This is in part because DRO is specifically designed for the distribution shift context; the synthetic control method does not come with specific robustness guarantees against adversarial distribution shift.

For CATE estimation, both methods perform equivalently well, with Optimal Transport methods weakly dominant.

This is largely because of the Nature of the CATE estimation task, in which the goal is to smoothly interpolate a function over a large covariate space. In this setting, the optimal site selection is a regularly spaced grid over the support of the covariates.

5.2.6 CATE methods perform poorly in the low-signal regime

Estimating the CATE is a fundamentally difficult problem, because it requires that we are able to well-estimate $\tau(x)$ at every 'cell' $X = x$. In the low-signal regime, our estimates will be inherently noisy.

The limited difference between CATE and PATE methods may be an artifact of the simulation structure.(Dehejia et al. 2019) argue that macro-level variables are, in the case they study, more significant moderators of treatment effects. By aggregating up individual level treatment effects, it is likely that we are constructing macro level variables with little realistic variation between sites, instead of supposing that treatment effects vary significantly as a function of macro variables.

When within-site variance of treatment effects is large relative to between-variance, selecting sites based on aggregate-level data is not very informative. This will naturally be the case when selecting sites based on aggregated data: we lose the individual-level information that ultimately determines how precise our estimate of the PEHE is.

In essence, even though we are in a high-signal regime, our site selection covariates are not especially predictive of individual treatment effects. We essentially need to study the behavior of the CATE method when treatment effects contain large, site-moderated effects.

6 Discussion and Conclusions

6.1 Practical Guidance for Applied Researchers

This paper provides a framework for choosing experimental sites based on observable covariates. To apply these methods in practice:

> **Step 1: Gather informative covariate data.** Before site selection, collect covariate data on all candidate sites. These should include characteristics you believe predict treatment effect heterogeneity based on theory, prior studies, or pilot evidence.
>
> **Step 2: Assess covariate informativeness.** Assess whether your covariates explain sufficient treatment effect variation. If observable covariates explain >50% of variation ($R^2 > 0.5$), optimization will outperform randomization. This can be evaluated through prior experiments, pre-treatment outcomes, or pilot studies. Otherwise, use stratified random sampling.
>
> **Step 3: Select an estimand.** If you need a single average treatment effect for policy decisions, optimize for the PATE using 1-Wasserstein distance. If you need to understand how effects vary across populations, optimize for the CATE using 2-Wasserstein distance.
>
> **Step 4: Adjust for distribution shift.** If your deployment population differs from observed sites, use the Jaccard radius procedure (Section 3.5) to select appropriate robustness levels. The automated procedure provides moderate, high, and maximum robustness options based on your data.
>
> **Step 5: *Ex post* balance testing.** After site selection, verify that selected sites adequately represent the population on observable characteristics. Check (i) covariate means and variances across selected versus non-selected sites, (ii) maximum distance from any population site to its nearest selected site, and (iii) the distribution of population mass assigned to each selected site under the optimal transport plan. For enhanced diagnostics, apply prognostic balance testing (Bicalho et al. 2022) using pre-treatment outcomes or predictions from auxiliary models to assess whether selected sites capture relevant predictive variation beyond observed covariates.

6.2 Summary of Findings

Distributionally-Robust Optimization methods hedge against realistic uncertainty in the deployment of field experiments.

The Crépon reanalysis demonstrates that distributionally robust optimization provides insurance against population misspecification at realistic uncertainty levels. DRO methods become preferred when deployment populations differ from candidate sites by margins exceeding 1.7 times observed empirical variation.

Use of optimization tools incentivizes allocating more resources to the planning stage.

A practitioner objection to these methods might be that collecting data *before* engaging in an RCT is expensive or difficult, and that large-scale, policy-relevant RCTs are already difficult enough. I argue however that pre-emption is better than cure: given the expense and scale of many modern RCTs, improving pre-execution data collection may significantly increase the efficiency of the actual experimental estimate, making it much less likely that the experiment will fail due to random features of the selected experimental population, rather than the absence of a treatment effect. The performance gains documented in our simulation suggest that optimization-based site selection can justify additional planning costs. PATE estimation improvements of $20 - 70\%$ should justify the upfront costs of additional scoping work for most large-scale RCTs.

Optimal transport-based site selection methods should be particularly useful for large scale experimental planning.

Optimal transport methods scale better than alternatives to large experimental design problems. The computational advantages become more pronounced as site pools and covariate dimensions increase, making these approaches particularly suitable for multi-country or multi-region experimental programs.

Convex hull approaches are likely less reliable in high dimensions

SPS relies on the idea that non-selected sites can be well-approximated by convex combinations of selected sites. Most of the probability mass concentrates near the boundary of the convex hull, making interior approximation unreliable. This also means that the convex hull approach is computationally more challenging in higher dimensions.

Optimization methods need good covariate information to be useful; otherwise, use randomization.

We found that optimization methods perform well against randomization when covariates were only moderately informative ($R^2 \in [.51, .19]$). Further, the worst-case performance of optimization methods significantly exceeded the 95^{th} percentile performance of randomization in our simulations. If planners are able to collect prognostic information, they could use this to run better-powered experiments, with guarantees on worst-case error.

Fundamental Limits and Knightian Uncertainty.

A fundamental challenge in site selection is that we typically observe only a subset $P \subset \mathcal{P}$ of the universe of potential experimental sites, and the gap between P and $\mathcal{P}$ represents a form of Knightian uncertainty (Knight 1921; Sunstein 2023). While our distributionally robust optimization methods provide insurance against distribution shifts within a Wasserstein ball of radius ρ around the observed data, choosing ρ itself requires confronting irreducible uncertainty about the nature of unobserved sites. This uncertainty differs qualitatively from the statistical risk we can quantify within P: we cannot assign probabilities to different ways $\mathcal{P}$ might differ from P without making untestable assumptions. For instance, if infrastructure constraints systematically exclude remote rural sites from P, we face true uncertainty—not merely risk—about how treatment effects might differ in these unobserved contexts. Our data-driven procedure for selecting ρ (Section 3.5) provides a pragmatic approach by benchmarking robustness levels against observed variation. This limitation is not specific to our methods but reflects an inherent constraint in experimental site selection: optimization and robustification can only operate within the bounds of what we observe. The practical implication is that expanding the set of feasible experimental sites P may be as important as optimally selecting from within it.

6.3 Future work

Neyman allocation

If we have information about individual covariates in a given site, it would be possible to incorporate this information into the site selection problem (Neyman 1934; Rosenman et al. 2022). Intuitively, for the PATE, we would want to minimize the within-variance of selected sites: this simply increases the error of our downstream estimate. But for the CATE, within-variance of selected sites is heterogeneity to be exploited downstream. In both cases, we could incorporate prior information about the informativeness of sites into the objective function of the minimization problem (Bertsimas et al. 2015).

Selecting individual units

We can adapt this method to select individuals to enroll in an experiment, not just sites. This is a topic of particular interest in experimental planning in industry settings, where user bases may be large, and understanding the behavior of specific market segments is of core interest (Arbour et al. 2021).

Optimal Transport methods are likely well-suited to this case, because, discussed above, they are well-suited to high-dimensional problems, and large-sized problem instances.

Optimal transport and DRO are applicable to a wide variety of core causal inference tasks.

It is possible to apply optimal transport methods to core tasks in causal inference: achieving balance between treatment and control distributions, matching, and synthetic control-type approaches. Distributionally Robust Optimization methods could be useful for researchers who want to assess the robustness of their conclusions to distribution shift. Here, the connection with sensitivity analysis is germane: researchers can find treatment effect estimates with guarantees on their stability under worst-case distribution shift.

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.

Amari, S. (2012). *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics. Springer New York. ISBN: 9781461250562. URL: https://books.google.com/books?id=XiDnBwAAQBAJ.

Ambrosio, L. (2003). *Optimal transport maps in Monge-Kantorovich problem*. arXiv: math/0304389 [math.AP]. URL: https://arxiv.org/abs/math/0304389.

Anderson, G. L., M. Limacher, A. R. Assaf, T. Bassford, S. A. Beresford, H. Black, D. Bonds, R. Brunner, R. Brzyski, B. Caan, et al. (2004). “Effects of Conjugated Equine Estrogen in Postmenopausal Women with Hysterectomy: The Women’s Health Initiative Randomized Controlled Trial”. In: *JAMA* 291.14, pp. 1701–1712. DOI: 10.1001/jama.291.14.1701.

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.

Arbour, D., D. Dimmery, and A. Rao (13–15 Apr 2021). “Efficient Balanced Treatment Assignments for Experimentation”. In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Ed. by A. Banerjee and K. Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, pp. 3070–3078. URL: https://proceedings.mlr.press/v130/arbour21a.html.

Athey, S. and G. W. Imbens (2006). “Identification and Inference in Nonlinear Difference-in-Differences Models”. In: *Econometrica* 74.2, pp. 431–497. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/3598807 (visited on 06/03/2022).

Banerjee, A., R. Banerji, J. Berry, E. Duflo, H. Kannan, S. Mukherji, M. Shotland, and M. Walton (2016). *Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of ”Teaching at the Right Level” in India*. Working Paper 22746. National Bureau of Economic Research.

— (2017). “From Proof of Concept to Scalable Policies: Challenges and Solutions, with an Application”. In: *Journal of Economic Perspectives* 31.4, pp. 73–102. DOI: 10.1257/jep.31.4.73.

Banerjee, A., S. Cole, E. Duflo, and L. Linden (2007). “Remedying Education: Evidence from Two Randomized Experiments in India”. In: *The Quarterly Journal of Economics* 122.3, pp. 1235–1264. DOI: 10.1162/qjec.122.3.1235.

Banerjee, A., E. Duflo, N. Goldberg, D. Karlan, R. Osei, W. Parienté, J. Shapiro, B. Thuysbaert, and C. Udry (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. Rothenhäusler (2023). *Learning under random distributional shifts*. arXiv: 2306.02948 [stat.ML]. URL: https://arxiv.org/abs/2306.02948.

Ben-Michael, E., A. Feller, D. A. Hirshberg, and J. R. Zubizarreta (2021). *The Balancing Act in Causal Inference*. arXiv: 2110.14831 [stat.ME]. URL: https://arxiv.org/abs/2110.14831.

Ben-Tal, A., D. den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen (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 (Aug. 2015). "The Power of Optimization Over Randomization in Designing Experiments Involving Small Samples". In: *Operations Research* 63.4, pp. 868–876. ISSN: 1526-5463. DOI: 10.1287/opre.2015.1361. URL: http://dx.doi.org/10.1287/opre.2015.1361.

Bertsimas, D. and M. Sim (Feb. 2004). "The Price of Robustness". In: *Operations Research* 52.1, pp. 35–53. ISSN: 1526-5463. DOI: 10.1287/opre.1030.0065. URL: http://dx.doi.org/10.1287/opre.1030.0065.

Bicalho, C., A. Bouyamourn, and T. Dunning (2022). *Conditional Balance Tests: Increasing Sensitivity and Specificity With Prognostic Covariates*. arXiv: 2205.10478 [stat.ME].

Blair, G., F. Christia, and J. M. Weinstein, eds. (2024). *Crime, Insecurity, and Community Policing: Experiments on Building Trust*. Studies in Comparative Politics. Cambridge University Press.

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., J. Li, S. Lin, and X. Zhang (2024). *Distributionally Robust Optimization and Robust Statistics*. arXiv: 2401.14655 [stat.ME]. URL: https://arxiv.org/abs/2401.14655.

Blanchet, J. and K. Murthy (2019a). "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. and N. Si (2019b). "Optimal uncertainty size in distributionally robust inverse covariance estimation". In: *Operations Research Letters* 47.6, pp. 618–621. ISSN: 0167-6377. DOI: https://doi.org/10.1016/j.orl.2019.10.005. URL: https://www.sciencedirect.com/science/article/pii/S0167637719300732.

Bloom, H. S., S. W. Raudenbush, M. J. Weiss, and K. P. and (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., M. Kimenyi, G. Mwabu, A. Ng'ang'a, and J. Sandefur (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.

Box, G. E. P. and K. B. Wilson (Dec. 1951). “On the Experimental Attainment of Optimum Conditions”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 13.1, pp. 1–38. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1951.tb00067.x. eprint: https://academic.oup.com/jrsssb/article-pdf/13/1/1/49093871/jrsssb_13_1_1.pdf. URL: https://doi.org/10.1111/j.2517-6161.1951.tb00067.x.

Box, G. and N. Draper (1987). *Empirical Model-Building and Response Surfaces*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780471810339. URL: https://books.google.com/books?id=QO2dDRufJEAC.

Box, G. E. P. and N. R. Draper (1975). “Robust Designs”. In: *Biometrika* 62.2, pp. 347–352. ISSN: 00063444. URL: http://www.jstor.org/stable/2335371 (visited on 06/02/2024).

Bradley, S., A. Hax, and T. Magnanti (1977). *Applied Mathematical Programming*. Addison-Wesley Publishing Company. ISBN: 9780201004649. URL: https://books.google.com/books?id=MSWdWv3Gn5cC.

Brown, G. W. (1951). “Iterative Solution of Games by Fictitious Play”. In: *Activity Analysis of Production and Allocation*. Ed. by T. C. Koopmans. New York: Wiley.

Cai, T. T., H. Namkoong, and S. Yadlowsky (2023). *Diagnosing Model Performance Under Distribution Shift*. arXiv: 2303.02011 [stat.ML]. URL: https://arxiv.org/abs/2303.02011.

Charpentier, A., E. Flachaire, and E. Gallic (2023). *Optimal Transport for Counterfactual Estimation: A Method for Causal Inference*. arXiv: 2301.07755 [econ.EM]. URL: https://arxiv.org/abs/2301.07755.

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.

Colmer, J., R. Martin, M. Muûls, and U. J. Wagner (2025). “Does Pricing Carbon Mitigate Climate Change? Firm-Level Evidence from the European Union Emissions Trading System”. In: *The Review of Economic Studies* 92.3, pp. 1625–1660. DOI: 10.1093/restud/rdae055.

Crépon, B., F. Devoto, E. Duflo, and W. Parienté (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.

Cuturi, M. (2013). “Sinkhorn Distances: Lightspeed Computation of Optimal Transport”. In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc., pp. 2292–2300.

Dawber, T. R., F. E. Moore, and G. V. Mann (1957). “Coronary Heart Disease in the Framingham Study”. In: *American Journal of Public Health and the Nation’s Health* 47.Supplement, pp. 4–24.

Dechezleprêtre, A., D. Nachtigall, and F. Venmans (2023). “The joint impact of the European Union emissions trading system on carbon emissions and economic performance”. In: *Journal of Environmental Economics and Management* 118, p. 102758. DOI: 10.1016/j.jeem.2022.102758.

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.

Deville, J.-C. and Y. Tillé (2004). “Efficient Balanced Sampling: The Cube Method”. In: *Biometrika* 91.4, pp. 893–912. ISSN: 00063444. URL: http://www.jstor.org/stable/20441151 (visited on 06/05/2025).

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.

Dunipace, E. (2022). *Optimal transport weights for causal inference.* arXiv: 2109.01991 [stat.ME]. URL: https://arxiv.org/abs/2109.01991.

Dunning, T., G. Grossman, M. Humphreys, S. D. Hyde, C. McIntosh, and G. Nellis, eds. (2019a). *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I.* Cambridge University Press.

Dunning, T. et al. (2019b). “Voter information campaigns and political accountability: Cumulative findings from a preregistered meta-analysis of coordinated trials”. In: *Science Advances* 5.7, eaaw2612. DOI: 10.1126/sciadv.aaw2612.

Ebersole, C. R. et al. (2020). “Many Labs 5: Testing pre-data-collection peer review as an intervention to increase replicability”. In: *Advances in Methods and Practices in Psychological Science* 3.

Ebersole, C. R., O. E. Atherton, A. L. Belanger, H. M. Skulborstad, J. M. Allen, J. B. Banks, E. Baranski, M. J. Bernstein, D. B. Bonfiglio, L. Boucher, et al. (2016). “Many Labs 3: Evaluating participant pool quality across the academic semester via replication”. In: *Journal of Experimental Social Psychology* 67, pp. 68–82. DOI: 10.1016/j.jesp.2015.10.012.

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.*

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.

Fournier, N. and A. Guillin (2013). *On the rate of convergence in Wasserstein distance of the empirical measure*. arXiv: 1312.2128 [math.PR]. URL: https://arxiv.org/abs/1312.2128.

Galichon, A. (2016). *Optimal Transport Methods in Economics*. Economics Books 10870. Princeton University Press. URL: https://ideas.repec.org/b/pup/pbooks/10870.html.

Gao, R., X. Chen, and A. J. Kleywegt (2020). *Wasserstein Distributionally Robust Optimization and Variation Regularization*. arXiv: 1712.06050 [cs.LG]. URL: https://arxiv.org/abs/1712.06050.

Gechter, M., K. Hirano, J. Lee, M. Mahmud, O. Mondal, J. Morduch, S. Ravindran, and A. S. Shonchoy (2024). *Selecting Experimental Sites for External Validity*. arXiv: 2405.13241 [econ.GN].

Gunsilius, F. F. (2025). *A primer on optimal transport for causal inference with observational data*. arXiv: 2503.07811 [stat.ME]. URL: https://arxiv.org/abs/2503.07811.

Hájek, J. (1971). "Contribution to discussion of paper by D. Basu". In: *Foundations of Statistical Inference*.

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.

Hill, J. L. (2011). "Bayesian Nonparametric Modeling for Causal Inference". In: *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240. DOI: 10.1198/jcgs.2010.08162. eprint: https://doi.org/10.1198/jcgs.2010.08162. URL: https://doi.org/10.1198/jcgs.2010.08162.

Horvitz, D. G. and D. J. Thompson (1952). "A Generalization of Sampling Without Replacement From a Finite Universe". In: *Journal of the American Statistical Association* 47.260, pp. 663–685. ISSN: 01621459. URL: http://www.jstor.org/stable/2280784 (visited on 09/14/2022).

Hu, Y., H. Zhu, E. Brunskill, and S. Wager (2024). *Minimax-Regret Sample Selection in Randomized Experiments*. arXiv: 2403.01386 [stat.ME]. URL: https://arxiv.org/abs/2403.01386.

Huang, M., N. Egami, E. Hartman, and L. Miratrix (Sept. 2023). "Leveraging population outcomes to improve the generalization of experimental results: Application to the JTPA study". In: *The Annals of Applied Statistics* 17.3. ISSN: 1932-6157. DOI: 10.1214/22-aoas1712. URL: http://dx.doi.org/10.1214/22-AOAS1712.

Hyde, S. D., E. Malesky, A. Coppock, M. Poertner, L. Young, et al. (2022). *Metaketa V: Women's Action Committees and Local Services*. DOI: 10.17605/OSF.IO/42PQ9.

Jin, Y., N. Egami, and D. Rothenhäusler (2024). *Beyond Reweighting: On the Predictive Role of Covariate Shift in Effect Generalization*. arXiv: 2412.08869 [stat.AP]. URL: https://arxiv.org/abs/2412.08869.

Jin, Y., K. Guo, and D. Rothenhausler (2023). "Diagnosing the role of observable distribution shift in scientific replications". In: URL: https://api.semanticscholar.org/CorpusID:261530919.

Kallus, N. (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.

Kannel, W. B., W. P. Castelli, P. M. McNamara, M. A. McKee, and M. Feinleib (1972). “Role of Blood Pressure in the Development of Congestive Heart Failure: The Framingham Study”. In: *New England Journal of Medicine* 287.16, pp. 781–787. DOI: 10.1056/NEJM197210192871601.

Kannel, W. B., T. R. Dawber, A. Kagan, N. Revotskie, and J. Stokes (1961). “Factors of Risk in the Development of Coronary Heart Disease — Six Year Follow-up Experience: The Framingham Study”. In: *Annals of Internal Medicine* 55, pp. 33–50.

Karmarkar, N. (Dec. 1984). “A New Polynomial-Time Algorithm for Linear Programming-II”. In: *Combinatorica* 4, pp. 373–395. DOI: 10.1007/BF02579150.

Klein, R. A., K. A. Ratliff, M. Vianello, R. B. Adams Jr, Š. Bahník, M. J. Bernstein, K. Bocian, M. J. Brandt, B. Brooks, C. C. Brumbaugh, et al. (2014). “Investigating variation in replicability: A ”Many Labs” replication project”. In: *Social Psychology* 45.3, pp. 142–152. DOI: 10.1027/1864-9335/a000178.

Klein, R. A., M. Vianello, F. Hasselman, B. G. Adams, R. B. Adams Jr, S. Alper, M. Aveyard, J. R. Axt, M. T. Babalola, Š. Bahník, et al. (2018). “Many Labs 2: Investigating Variation in Replicability Across Samples and Settings”. In: *Advances in Methods and Practices in Psychological Science* 1.4, pp. 443–490. DOI: 10.1177/2515245918810225.

Klemetsen, M. E., K. E. Rosendahl, and A. L. Jakobsen (2020). “The impacts of the EU ETS on Norwegian plants’ environmental and economic performance”. In: *Climate Change Economics* 11.1, pp. 1–32. DOI: 10.1142/S2010007820500062.

Knight, F. H. (1921). *Risk, Uncertainty and Profit.* Boston, MA: Houghton Mifflin Co. URL: http://www.econlib.org/library/Knight/knRUP.html.

Kuhn, D., P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh (2024). *Wasserstein Distributionally Robust Optimization: Theory and Applications in Machine Learning.* arXiv: 1908.08729 [stat.ML]. URL: https://arxiv.org/abs/1908.08729.

Levy, D., Y. Carmon, J. C. Duchi, and A. Sidford (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.

Manson, J. E., R. T. Chlebowski, M. L. Stefanick, A. K. Aragaki, J. E. Rossouw, R. L. Prentice, G. Anderson, B. V. Howard, C. A. Thomson, A. Z. LaCroix, et al. (2013). “Menopausal Hormone Therapy and Health Outcomes During the Intervention and Extended Poststopping Phases of the Women’s Health Initiative Randomized Trials”. In: *JAMA* 310.13, pp. 1353–1368. DOI: 10.1001/jama.2013.278040.

Manson, J. E., C. J. Crandall, J. E. Rossouw, R. T. Chlebowski, G. L. Anderson, M. L. Stefanick, A. K. Aragaki, J. A. Cauley, G. L. Wells, A. Z. LaCroix, et al. (2024). “The Women’s Health Initiative Randomized Trials and Clinical Practice: A Review”. In: *JAMA* 331.20, pp. 1748–1760. DOI: 10.1001/jama.2024.6542.

Monge, G. (1781). *Mémoire sur la théorie des déblais et des remblais.* Memoirs de l’Académie Royale des Sciences de Paris, pp. 666–704. Paris: De l’Imprimerie Royale.

Murray, M. K. and J. W. Rice (Apr. 1993). *Differential geometry and statistics.* en. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. Philadelphia, PA: Chapman & Hall/CRC.

Natarajan, B. K. (1995). “Sparse Approximate Solutions to Linear Systems”. In: *SIAM J. Comput.* 24, pp. 227–234. URL: https://api.semanticscholar.org/CorpusID:2072045.

Neyman, J. (Dec. 1934). “On the Two Different Aspects of the Representative Method : The Method of Stratified Sampling and the Method of Purposive Selection”. In: *Journal of the Royal Statistical Society* 97.4, pp. 558–606. ISSN: 0952-8385. DOI: 10.1111/j.2397-2335.1934.tb04184.x. eprint: https://academic.oup.com/jrsssa/article-pdf/97/4/558/49706846/jrsssa_97_4_558.pdf. URL: https://doi.org/10.1111/j.2397-2335.1934.tb04184.x.

Olea, J. L. M., B. Prallon, C. Qiu, J. Stoye, and Y. Sun (2024). *Externally Valid Selection of Experimental Sites via the k-Median Problem.* arXiv: 2408.09187 [econ.EM]. URL: https://arxiv.org/abs/2408.09187.

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.

Rosenman, E. T. R. and L. Miratrix (2022). *Designing Experiments Toward Shrinkage Estimation.* arXiv: 2204.06687 [stat.ME]. URL: https://arxiv.org/abs/2204.06687.

Rossouw, J. E., G. L. Anderson, R. L. Prentice, A. Z. LaCroix, C. Kooperberg, M. L. Stefanick, R. D. Jackson, S. A. Beresford, B. V. Howard, K. C. Johnson, 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.

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

Roughgarden, T. (Aug. 2016). *Twenty Lectures on Algorithmic Game Theory.* Cambridge University Press. ISBN: 9781316779309. DOI: 10.1017/cbo9781316779309. URL: http://dx.doi.org/10.1017/CBO9781316779309.

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., N. T. Williams, E. A. Stuart, and I. Diaz (2024). *Improving efficiency in transporting average treatment effects.* arXiv: 2304.00117 [stat.ME]. URL: https://arxiv.org/abs/2304.00117.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling.* Progress in Nonlinear Differential Equations and Their Applications. Cham: Springer.

Saville, B. R., D. A. Berry, N. S. Berry, K. Viele, and S. M. Berry (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.

Shalit, U., F. D. Johansson, and D. Sontag (2017). *Estimating individual treatment effect: generalization bounds and algorithms*. arXiv: 1606.03976 [stat.ML]. URL: https://arxiv.org/abs/1606.03976.

Shpitser, I., T. VanderWeele, and J. M. Robins (2012). *On the Validity of Covariate Adjustment for Estimating Causal Effects*. arXiv: 1203.3515 [stat.ME]. URL: https://arxiv.org/abs/1203.3515.

Slough, T. et al. (2021). “Adoption of community monitoring improves common pool resource management across contexts”. In: *Proceedings of the National Academy of Sciences* 118.29, e2015367118. DOI: 10.1073/pnas.2015367118.

Stuart, E. A., B. K. Lee, and F. P. Leacy (Aug. 2013). “Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research”. In: *Journal of Clinical Epidemiology* 66.8, S84–S90.e1. ISSN: 0895-4356. DOI: 10.1016/j.jclinepi.2013.01.013. URL: http://dx.doi.org/10.1016/j.jclinepi.2013.01.013.

Sunstein, C. R. (2023). “Knightian Uncertainty”. In: *SSRN Electronic Journal*. ISSN: 1556-5068. DOI: 10.2139/ssrn.4662711. URL: http://dx.doi.org/10.2139/ssrn.4662711.

Tan, Y. Y., V. Papez, W. H. Chang, S. H. Mueller, S. Denaxas, and A. G. Lai (Oct. 2022). “Comparing clinical trial population representativeness to real-world populations: an external validity analysis encompassing 43 895 trials and 5 685 738 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.

Taori, R., A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt (2020). *Measuring Robustness to Natural Distribution Shifts in Image Classification*. arXiv: 2007.00644 [cs.LG]. URL: https://arxiv.org/abs/2007.00644.

Thompson, R. (May 2022). “Robust subset selection”. In: *Computational Statistics amp; Data Analysis* 169, p. 107415. ISSN: 0167-9473. DOI: 10.1016/j.csda.2021.107415. URL: http://dx.doi.org/10.1016/j.csda.2021.107415.

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.

Torous, W., F. Gunsilius, and P. Rigollet (2024). *An Optimal Transport Approach to Estimating Causal Effects via Nonlinear Difference-in-Differences*. arXiv: 2108.05858 [stat.ME]. URL: https://arxiv.org/abs/2108.05858.

VanderWeele, T. J. and I. Shpitser (May 2011). “A New Criterion for Confounder Selection”. In: *Biometrics* 67.4, pp. 1406–1413. ISSN: 0006-341X. DOI: 10.1111/j.1541-0420.2011.01619.x. URL: http://dx.doi.org/10.1111/j.1541-0420.2011.01619.x.

Vershik, A. M. (May 2013). “Long History of the Monge-Kantorovich Transportation Problem: (Marking the centennial of L.V. Kantorovich’s birth!)” In: *The Mathematical Intelligencer* 35.4, pp. 1–9. ISSN: 1866-7414. DOI: 10.1007/s00283-013-9380-x. URL: http://dx.doi.org/10.1007/s00283-013-9380-x.

Villani, C. and A. M. Society (2003). *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society. ISBN: 9781470418045. URL: https://books.google.com/books?id=MyPjjgEACAAJ.

Villani, C. (Mar. 2003). *Topics in Optimal Transportation*. American Mathematical Society. ISBN: 9781470418045. DOI: 10.1090/gsm/058. URL: http://dx.doi.org/10.1090/gsm/058.

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

WHO Solidarity Trial Consortium (2022). “Remdesivir and three other drugs for hospitalised patients with COVID-19: final results of the WHO Solidarity randomised trial and updated meta-analyses”. In: *The Lancet* 399.10339, pp. 1941–1953. DOI: 10.1016/S0140-6736(22)00519-0.

WHO Solidarity Trial Consortium, H. Pan, R. Peto, A.-M. Henao-Restrepo, M.-P. Preziosi, V. Sathiyamoorthy, Q. Abdool Karim, M. M. Alejandria, C. Hernández García, M.-P. Kieny, 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.

Wu, C.-F. (Nov. 1981). “On the Robustness and Efficiency of Some Randomized Designs”. In: *The Annals of Statistics* 9.6. ISSN: 0090-5364. DOI: 10.1214/aos/1176345634. URL: http://dx.doi.org/10.1214/aos/1176345634.

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.

A Proofs of Main Results

A.1 Technical Preliminaries

We first show three minor results that are needed to state the proof of the two main theorems in the text.

Lemma 22 (Corollary of Kantorovich-Rubenstein Formula). *If f is Lipschitz, then*

$$\left|\int f\,d\mu - \int f\,d\nu\right| \leq L \cdot W_1(\mu,\nu)$$

Proof. The Kantorovich-Rubinstein Formula states: If f is Lipschitz with constant L, then:

$$\begin{aligned}\left|\int f\,d\mu - \int f\,d\nu\right| &\leq \sup_h \left\{\left|\int h\,d\mu - \int h\,d\nu\right| : h \text{ is 1-Lipschitz}\right\}\\ &= W_1(\mu,\nu)\end{aligned}$$

Define $g(x) = \frac{f(x)}{L}$. Then:

$$\begin{aligned}\left|\int g\,d\mu - \int g\,d\nu\right| &\leq W_1(\mu,\nu)\\ \left|\int \frac{f}{L}\,d\mu - \int \frac{f}{L}\,d\nu\right| &\leq W_1(\mu,\nu)\\ \frac{1}{L}\left|\int f\,d\mu - \int f\,d\nu\right| &\leq W_1(\mu,\nu)\\ \left|\int f\,d\mu - \int f\,d\nu\right| &\leq L \cdot W_1(\mu,\nu)\end{aligned}$$

□

We also require two facts about Wasserstein Distances:

Lemma 23 (Wasserstein Distance with Shared Conditionals). *If $P_{X,U} = P_X \times P_{U|X}$ and $Q_{X,U} = Q_X \times P_{U|X}$ are two joint distributions that share the same conditional distribution $P_{U|X}$ but have different marginals P_X and Q_X, then:*

$$W_p(P_{X,U}, Q_{X,U}) = W_p(P_X, Q_X)$$

Proof. To show that $W_p(P_{X,U}, Q_{X,U}) = W_p(P_X, Q_X)$, we need to show that $W_p(P_{X,U}, Q_{X,U}) \leq W_p(P_X, Q_X)$ and $W_p(P_X, Q_X) \leq W_p(P_{X,U}, Q_{X,U})$. First, we show that $W_p(P_{X,U}, Q_{X,U}) \leq W_p(P_X, Q_X)$.

Let γ_X^* be an optimal transport plan between P_X and Q_X, so that:

$$\int |x_1 - x_2|^p \, d\gamma_X^*(x_1, x_2) = W_p^p(P_X, Q_X)$$

We define a transport plan Π^* for $P_{X,U}$ and $Q_{X,U}$ by setting:

$$d\Pi^*((x_1,u_1),(x_2,u_2)) = d\gamma_X^*(x_1,x_2)K(du_1|x_1)\delta_{u_1}(du_2)$$

Where $\delta_{u_1}(du_2)$ implies $u_2 = u_1$. The first marginal of Π^* is:

$$\int_{x_2,u_2} d\Pi^*((x_1,u_1),(x_2,u_2)) = K(du_1|x_1)\int_{x_2} d\gamma_X^*(x_1,x_2) = K(du_1|x_1)dP_X(x_1) = dP_{X,U}(x_1,u_1)$$

The second marginal of Π^* is:

$$\int_{x_1,u_1} d\Pi^*((x_1,u_1),(x_2,u_2)) = \int_{x_1} K(du_2|x_1)d\gamma_X^*(x_1,x_2)$$

We can apply the Disintegration Theorem (see (Villani 2008), to show that, for shared kernel K and optimal γ_X^*, the second marginal can be written as:

$$dQ_X(x_2)K(du_2|x_2) = dQ_{X,U}(x_2,u_2)$$

.

The cost of Π^* is

$$C(\Pi^*) = \int (|x_1 - x_2|^p + |u_1 - u_2|^p)d\Pi^*$$

$u_1 = u_2$ by construction, so that $|u_1 - u_2|^p = 0$, giving us:

$$C(\Pi^*) = \int |x_1 - x_2|^p d\gamma_X^*(x_1,x_2)\left(\int K(du_1|x_1)\right)$$

.

Since $\int K(du_1|x_1) = 1$:

$$C(\Pi^*) = \int |x_1 - x_2|^p d\gamma_X^*(x_1,x_2) = W_p^p(P_X,Q_X)$$

Since $W_p^p(P_{X,U},Q_{X,U})$ is the infimal cost,

$$W_p^p(P_{X,U},Q_{X,U}) \leq C(\Pi^*) = W_p^p(P_X,Q_X)$$

Finally, because the p-Wasserstein distance is the p-th *root* of the optimal cost,

$$W_p(P_{X,U},Q_{X,U}) \;=\; \left(\inf_\gamma \int d((x,u),(x',u'))^p\,d\gamma\right)^{1/p} \;\leq\; \left(\int |x_1 - x_2|^p\,d\gamma_X^*\right)^{1/p} \;=\; W_p(P_X,Q_X).$$

This entails that:

$$W_p(P_{X,U},Q_{X,U}) \leq W_p(P_X,Q_X)$$

As required.

For the reverse direction, consider any transport plan γ between $P_{X,U}$ and $Q_{X,U}$. Define:

$$\gamma_X(x_1,x_2) = \int_{u_1}\int_{u_2} \gamma((x_1,u_1),(x_2,u_2))\,du_2\,du_1$$

This gives a transport plan between P_X and Q_X. The cost of this plan is less than or equal to the cost of γ:

$$\int_{x_1,x_2} |x_1 - x_2|^p\,d\gamma_X(x_1,x_2) \leq \iint (|x_1 - x_2|^p + |u_1 - u_2|^p)\,d\gamma((x_1,u_1),(x_2,u_2))$$

Since $W_p^p(P_X, Q_X)$ is the minimum cost over all transport plans between P_X and Q_X:

$$W_p^p(P_X, Q_X) \leq \int_{x_1,x_2} |x_1 - x_2|^p \, d\gamma_X(x_1, x_2) \leq C(\gamma)$$

Taking the p^{th} root, we have:

$$W_p(P_X, Q_X) \leq \left(\int_{x_1,x_2} |x_1 - x_2|^p \, d\gamma_X(x_1, x_2) \right)^{\frac{1}{p}} \leq \left(\inf_{\gamma} \int d((x,u),(x',u'))^p \, d\gamma \right)^{1/p} = W_p(P_{X,U}, Q_{X,U})$$

This implies $W_p(P_X, Q_X) \leq W_p(P_{X,U}, Q_{X,U})$.

Combining the two inequalities, we have:

$$W_p(P_{X,U}, Q_{X,U}) = W_p(P_X, Q_X)$$

□

Lemma 24 (Wasserstein Distance with Shared Marginals). *If $P_{X,U} = F_X \times P_{U|X}$ and $Q_{X,U} = F_X \times Q_{U|X}$ are two joint distributions with the same marginal distribution F_X but different conditional distributions $P_{U|X}$ and $Q_{U|X}$, then:*

$$W_p(P_{X,U}, Q_{X,U}) = \int W_p(P_{U|X=x}, Q_{U|X=x}) \, dF_X(x) = \mathbb{E}_{F_X}[W_p(P_{U|X}, Q_{U|X})]$$

Proof. We will show that the optimal transport plan works independently within each slice corresponding to a specific value of $X = x$.

For any joint distribution γ on $(X \times U) \times (X \times U)$ with marginals $P_{X,U}$ and $Q_{X,U}$, define:

$$\gamma_X(x_1, x_2) = \int_{u_1} \int_{u_2} \gamma((x_1, u_1), (x_2, u_2)) \, du_2 \, du_1$$

Since both $P_{X,U}$ and $Q_{X,U}$ have the same marginal F_X, any transport plan γ with these marginals must have:

$$\gamma_X(x_1, x_2) = \begin{cases} F_X(x_1) & \text{if } x_1 = x_2 \\ 0 & \text{if } x_1 \neq x_2 \end{cases}$$

This means $\gamma((x_1, u_1), (x_2, u_2)) = 0$ whenever $x_1 \neq x_2$.

We can express any transport plan γ as:

$$\gamma((x, u_1), (x, u_2)) = F_X(x) \cdot \gamma_x(u_1, u_2)$$

where for each x, γ_x is a transport plan between $P_{U|X=x}$ and $Q_{U|X=x}$.

The total transportation cost is:

$$\begin{aligned} C(\gamma) &= \iint d((x_1, u_1), (x_2, u_2))^p \, d\gamma((x_1, u_1), (x_2, u_2)) \\ &= \iint (|x_1 - x_2|^p + |u_1 - u_2|^p) \, d\gamma((x_1, u_1), (x_2, u_2)) \end{aligned}$$

Since γ only assigns probability to pairs where $x_1 = x_2 = x$, and $|x - x|^p = 0$:

$$\begin{aligned} C(\gamma) &= \iint |u_1 - u_2|^p \, d\gamma((x, u_1), (x, u_2)) \\ &= \int_x F_X(x) \left(\iint |u_1 - u_2|^p \, d\gamma_x(u_1, u_2) \right) \, dx \end{aligned}$$

For each x, the minimum value of $\iint |u_1 - u_2|^p \, d\gamma_x(u_1, u_2)$ is exactly $W_p^p(P_{U|X=x}, Q_{U|X=x})$ by the definition of the Wasserstein distance.

Therefore, the minimum total cost is:

$$\begin{aligned} W_p^p(P_{X,U}, Q_{X,U}) &= \int F_X(x) \cdot W_p^p(P_{U|X=x}, Q_{U|X=x}) \, dx \\ &= \int W_p^p(P_{U|X=x}, Q_{U|X=x}) \, dF_X(x) \end{aligned}$$

Taking the p-th root:

$$W_p(P_{X,U}, Q_{X,U}) = \left(\int W_p^p(P_{U|X=x}, Q_{U|X=x}) \, dF_X(x) \right)^{1/p}$$

□

Corollary 25. *For $p = 1$, we have:*

$$\begin{aligned} W_1(P_{X,U}, Q_{X,U}) &= \int W_1(P_{U|X=x}, Q_{U|X=x}) \, dF_X(x) \\ &= \mathbb{E}_{F_X}[W_1(P_{U|X}, Q_{U|X})] \end{aligned}$$

A.2 Proof of Theorem 14

Theorem 26 (Upper Bound on PATE MSE)**.** *Under the stated assumptions, the Mean Squared Error of the PATE estimator is bounded by:*

$$MSE_{PATE} \leq L^2 \cdot (W_1(P_X, S_X) + \eta)^2 + \sigma_S^2$$

where $\eta = \mathbb{E}_{P_X}[W_1(P_{U|X}, S_{U|X})]$ represents the degree of unmeasured heterogeneity, and σ_S^2 is the error of the downstream treatment effect estimator.

Proof. Starting with the definition of MSE_{PATE}, we have:

$$\begin{aligned} \text{MSE}_{PATE} &= \mathbb{E}\left[\left(\tau^P - \hat{\tau}^S \right)^2 \right] \\ &= \left(\int \tau(x, u) \, dF_P(x, u) - \int \hat{\tau}(x, u) \, dF_S(x, u) \right)^2 \\ &= \left(\int \tau(x, u) \, dF_P(x, u) - \int \tau(x, u) \, dF_S(x, u) + \int \tau(x, u) \, dF_S(x, u) - \int \hat{\tau}(x, u) \, dF_S(x, u) \right)^2 \\ &= \left(\int \tau(x, u)[dF_P(x, u) - dF_S(x, u)] + \int [\tau(x, u) - \hat{\tau}(x, u)] dF_S(x, u) \right)^2 \end{aligned}$$

By Assumption 10 (independence of treatment assignment and site selection):

$$\text{MSE}_{PATE} = \left(\int \tau(x,u)[dF_P(x,u) - dF_S(x,u)]\right)^2 + \left(\int [\tau(x,u) - \hat{\tau}(x,u)]dF_S(x,u)\right)^2$$

Define $\sigma_S^2 = \left(\int [\tau(x,u) - \hat{\tau}(x,u)]dF_S(x,u)\right)^2$, which is the sampling error of our estimator of τ. From the perspective of our argument, this is irreducible noise.

This gives us:

$$\text{MSE}_{PATE} = \left(\int \tau(x,u)[dF_P(x,u) - dF_S(x,u)]\right)^2 + \sigma_S^2$$

Now, since, by Assumption 9, $\tau(x,u)$ is Lipschitz with constant L, we can apply Lemma 22 to get an upper bound on the error due to difference in distributions P and S:

$$\left(\int \tau(x,u)[dF_P(x,u) - dF_S(x,u)]\right)^2 \leq L^2 \cdot W_1^2(P_{X,U}, S_{X,U})$$

We can now decompose the joint Wasserstein distance between $P_{X,U}$ and $S_{X,U}$ into components related to the observed covariates X and unobserved covariates U.

First, define $Q_{X,U} = P_X \times S_{U|X}$, which has the marginal distribution of X from the population (P_X) but the conditional distribution of U given X from the selected sites $(S_{U|X})$. Then, since the Wasserstein distance is a proper metric, we can apply the triangle inequality, so that:

$$W_1(P_{X,U}, S_{X,U}) \leq W_1(P_{X,U}, Q_{X,U}) + W_1(Q_{X,U}, S_{X,U})$$

Consider the terms on the right hand side. First, by Lemma 23, we have that:

$$W_1(Q_{X,U}, S_{X,U}) = W_1(P_X, S_X)$$

And by Lemma 24, we have that:

$$W_1(P_{X,U}, Q_{X,U}) = \int W_1(P_{U|X}, S_{U|X})\, dF_{P_X} = \mathbb{E}_{P_X}\left[W_1(P_{U|X}, S_{U|X})\right]$$

So that:

$$W_1(P_{X,U}, S_{X,U}) \leq \mathbb{E}_{P_X}\left[W_1(P_{U|X}, S_{U|X})\right] + W_1(P_X, S_X)$$

Consistent with practice in sensitivity analysis, let us reparameterize this quantity as follows:

$$\eta_1 \equiv \mathbb{E}_{P_X}\left[W_1(P_{U|X}, S_{U|X})\right]$$

Finally, we can return to upper bounding the MSE_{PATE}. We have:

$$\left(\int \tau(x,u)[dF_P(x,u) - dF_S(x,u)]\right)^2 \leq L^2 \cdot W_1^2(P_{X,U}, S_{X,U}) \leq L^2 \cdot [W_1(P_X, S_X) + \eta_1]^2$$

Putting this all together, we have:

$$MSE_{\text{PATE}} \leq L^2 \cdot [W_1(P_X, S_X) + \eta_1]^2 + \sigma_S^2$$

□

A.3 Proof of Theorem 15

Theorem 27 (Upper Bound on PEHE)**.** *Under the stated assumptions, the Precision in Estimation of Heterogeneous Effect is bounded by:*

$$PEHE \leq L^2 \cdot [W_2(P_X, S_X) + \eta_2]^2 + \sigma_S^2$$

where $\eta_2 = \mathbb{E}_{P_X}[W_2(P_{U|X}, S_{U|X})]$ *represents the effect of unmeasured heterogeneity, and* σ_S^2 *represents irreducible estimation error.*

Proof. Since treatment effects depend on both observed covariates x and unobserved covariates u, we work with the full covariate vector $\xi = (x, u)$ and treatment effects $\tau(\xi) = \tau(x, u)$. The PEHE can be written as:

$$\text{PEHE} = \iint [\tau^P(x,u) - \hat{\tau}^S(x,u)]^2 dP_{X,U}(x,u)$$

Using the decomposition $\tau^P(x,u) - \hat{\tau}^S(x,u) = [\tau^P(x,u) - \tau^S(x,u)] + [\tau^S(x,u) - \hat{\tau}^S(x,u)]$ and applying Assumption 10 (independence of experimental design and site selection):

$$\text{PEHE} = \iint [\tau^P(x,u) - \tau^S(x,u)]^2 dP_{X,U}(x,u) + \iint [\tau^S(x,u) - \hat{\tau}^S(x,u)]^2 dP_{X,U}(x,u)$$

Define the second term as the irreducible estimation error:

$$\sigma_S^2 = \iint [\tau^S(x,u) - \hat{\tau}^S(x,u)]^2 dP_{X,U}(x,u)$$

For the first term, we define $\tau^S(x,u)$ via the optimal transport plan π^* from $P_{X,U}$ to $S_{X,U}$:

$$\tau^S(x,u) = \iint \tau(x', u')\pi^*((x,u), d(x', u'))$$

By Assumption 9 (τ is L-Lipschitz):

$$\begin{aligned} |\tau^P(x,u) - \tau^S(x,u)| &= \left|\tau(x,u) - \iint \tau(x',u')\pi^*((x,u), d(x',u'))\right| \\ &\leq L \iint ||(x,u) - (x',u')||\pi^*((x,u), d(x',u')) \end{aligned}$$

Squaring both sides:

$$[\tau^P(x,u) - \tau^S(x,u)]^2 \leq L^2 \left[\iint ||(x,u) - (x',u')||\pi^*((x,u), d(x',u'))\right]^2$$

Since $\iint \pi^*((x,u), d(x',u')) = 1$, we apply Jensen's inequality:

$$\left[\iint ||(x,u) - (x',u')||\pi^*((x,u), d(x',u'))\right]^2 \leq \iint ||(x,u) - (x',u')||^2\pi^*((x,u), d(x',u'))$$

Therefore:

$$[\tau^P(x,u) - \tau^S(x,u)]^2 \leq L^2 \iint ||(x,u) - (x',u')||^2\pi^*((x,u), d(x',u'))$$

Integrating over $P_{X,U}$ and taking the infimum over all transport plans:

$$\iint [\tau^P(x,u) - \tau^S(x,u)]^2 dP_{X,U}(x,u) \leq L^2 W_2^2(P_{X,U}, S_{X,U})$$

Now we decompose the joint Wasserstein distance. Define $Q_{X,U} = P_X \times S_{U|X}$ and apply the triangle inequality:

$$W_2(P_{X,U}, S_{X,U}) \leq W_2(P_{X,U}, Q_{X,U}) + W_2(Q_{X,U}, S_{X,U})$$

By Proposition 23 (shared marginals):

$$W_2(P_{X,U}, Q_{X,U}) = \mathbb{E}_{P_X}[W_2(P_{U|X}, S_{U|X})] = \eta_2$$

By Proposition 22 (shared conditionals):

$$W_2(Q_{X,U}, S_{X,U}) = W_2(P_X, S_X)$$

Therefore:

$$W_2(P_{X,U}, S_{X,U}) \leq \eta_2 + W_2(P_X, S_X)$$

Substituting back:

$$\text{PEHE} \leq L^2 W_2^2(P_{X,U}, S_{X,U}) + \sigma_S^2 \leq L^2[\eta_2 + W_2(P_X, S_X)]^2 + \sigma_S^2$$

Rearranging:

$$\text{PEHE} \leq L^2[W_2(P_X, S_X) + \eta_2]^2 + \sigma_S^2$$

This completes the proof. □

A.4 Proof of Proposition 16

Proof. The goal is to minimize the p-Wasserstein distance $W_p(P_X, S_X)$ between the empirical distribution of covariates in the population (P_X) and the empirical distribution in the selected sites (S_X). We show that this minimization is equivalent to our mixed integer linear program.

The p-Wasserstein distance is defined:

$$W_p(P_X, S_X) = \left(\inf_{\gamma \in \Gamma(P_X, S_X)} \int \|x - y\|^p d\gamma(x,y) \right)^{1/p}$$

where $\Gamma(P_X, S_X)$ is the set of all joint distributions (transport plans) with marginals P_X and S_X.

For discrete distributions with finite support, this becomes:

$$W_p(P_X, S_X) = \left(\min_{\pi \in \Pi(P_X, S_X)} \sum_{i,j} \pi_{ij} \|x_i - x_j\|^p \right)^{1/p}$$

where π_{jk} represents the amount of probability mass transported from location x_i in the population to location x_j in the selected sites. Since the $(1/p)$-th power function is monotonically increasing, minimizing $W_p(P_X, S_X)$ is equivalent to minimizing $\sum_{i,j} \pi_{ij} \|x_i - x_j\|^p$.

The constraints arise from the site selection problem structure. The empirical distribution P_X assigns equal probability mass $\frac{1}{|P|}$ to each site in the population, yielding:

$$\sum_{k=1}^{|P|} \pi_{ij} = \frac{1}{|P|} \quad \forall i \in P$$

The empirical distribution S_X depends on the selection variables s_i, assigning mass:

$$S_X(x_i) = \begin{cases} \frac{1}{K} & \text{if site } i \text{ is selected } (s_i = 1) \\ 0 & \text{otherwise} \end{cases}$$

where $K = \sum_{j=1}^{|P|} s_i$ is the number of selected sites. This gives:

$$\sum_{j=1}^{|P|} \pi_{ij} = \frac{s_i}{\sum_{l=1}^{|P|} s_l} \quad \forall i \in P$$

We can only transport probability mass to selected sites: $\pi_{ij} \leq s_i$ for all $i, j \in P$. The site selection budget constraint limits us to at most K sites: $\sum_{i=1}^{|P|} s_i \leq K$. All transport plan entries must be non-negative: $\pi_{ij} \geq 0$ for all $i, j \in P$.

The objective function $\sum_{i=1}^{|P|} \sum_{j=1}^{|P|} \pi_{ij} \|x_i - x_j\|^p$ directly computes the p-Wasserstein distance (up to the monotonic transformation) given a valid transport plan. Therefore, minimizing $W_p(P_X, S_X)$ subject to selecting at most K sites is equivalent to solving the stated MILP. □

A.5 Proof of Proposition 20

Proof. We have $\text{UB}^{(t+1)} = W_p(Q^{(t+1)}, S^{(t)})$, which is Nature's best response to the current site selection. It is an upper bound because the optimal site selection S^* must minimize the worst-case distance, so it must perform at least as well as any feasible solution against Nature's worst-case attack:

$$\text{OPT} = \max_{Q: W_p(Q, P_X) \leq \rho} W_p(Q, S^*) \leq \max_{Q: W_p(Q, P_X) \leq \rho} W_p(Q, S^{(t)}) = \text{UB}^{(t+1)}$$

Likewise, $\text{LB}^{(t+1)} = \max_{Q \in \mathcal{Q}^{(t+1)}} W_p(Q, S^{(t+1)})$ is the Researcher's best response against all observed scenarios up to time t. This provides a lower bound because $S^{(t+1)}$ is the optimal solution to a relaxed version of the original problem:

$$\text{LB}^{(t+1)} = \min_{S: |S| = K} \max_{Q \in \mathcal{Q}^{(t+1)}} W_p(Q, S)$$

Since we only consider scenarios in $\mathcal{Q}^{(t+1)}$ rather than all possible adversarial distributions, the relaxed problem is easier than the original:

$$\mathcal{Q}^{(t+1)} \subseteq \{Q : W_p(Q, P_X) \leq \rho\}$$

Therefore, the optimal value of the relaxed problem provides a lower bound on the original problem:

$$\mathrm{LB}^{(t+1)} = \min_{S:|S|=K} \max_{Q\in\mathcal{Q}^{(t+1)}} W_p(Q, S^{(t+1)}) \leq \min_{S:|S|=K} \max_{Q:W_p(Q,P_X)\leq\rho} W_p(Q,S) = \max_{Q:W_p(Q,P_X)\leq\rho} W_p(Q,S^*) = \mathrm{OPT}$$

Combining these inequalities, we have:

$$\mathrm{LB}^{(t+1)} \leq \mathrm{OPT} \leq \mathrm{UB}^{(t+1)}$$

Recall that the algorithim terminates when $\mathrm{UB}^{(t+1)} - \mathrm{LB}^{(t+1)} < \epsilon$. But by the above, this implies that we have bracketed the true optimal value within ϵ, guaranteeing that $S^{(t+1)}$ is ϵ-close to S^*, as desired. □

B Simulation Details

B.1 Randomization versus Optimization

Simulation Design: We generate candidate populations of $S = 30$ sites with covariates $X_s \sim \mathcal{N}(0, I_5)$ and site-level treatment effects

$$U_s = \sqrt{1-\eta^2} f(X_s) + \eta\varepsilon_s, \quad \varepsilon_s \sim \mathcal{N}(0,1) \quad \tau_{is} = \beta^\top X_s + \gamma U_s + \xi_{is}, \quad \xi_{is} \sim \mathcal{N}(0,\sigma^2)$$

Parameter $\eta \in \{0, 0.25, 0.5, 0.75, 1\}$ controls the fraction of treatment heterogeneity unexplained by observed covariates: $\eta = 0$ implies all variation is explained ($U_s = f(X_s)$), while $\eta = 1$ implies purely idiosyncratic effects ($U_s = \varepsilon_s$). The population CATE is $\tau(x) = \beta^\top x + \gamma\sqrt{1-\eta^2} f(x)$ with PATE $\overline{\tau}^{\text{pop}} = \mathbb{E}[\tau_{is}]$.

Site Selection Methods: From each population, select K sites using:

Wasserstein Methods: OPT-PATE, OPT-CATE, DRO variants

Random Sampling: Uniform selection across sites

Stratified Sampling: k-means clustering + within-stratum sampling

Stochastic methods use $B = 500$ draws.

Evaluation: Fit CATE model $\widehat{\tau}^{(m,r,b)}(x)$ on selected sites and compute:

PATE: $\mathrm{MSE}_{\text{PATE}} = \left(\overline{\tau}^{\text{pop}} - \overline{\tau}^{(m,r,b)}\right)^2$

CATE: $\mathrm{PEHE} = \mathbb{E}\left[\tau_{is} - \widehat{\tau}^{(m,r,b)}(X_{is})\right]^2$

Where PEHE expectation is over all SN units. Average stochastic methods over B draws, then pool across $R = 10$ replications to report performance versus η.

Output: Performance comparison across 5 signal strength levels, evaluating optimization versus randomization trade-offs under varying treatment effect predictability.

B.2 Crépon et al.

Data Setup: Load Crépon et al. Morocco microcredit data. Generate 250 base datasets by sampling $|P| \in \{20, 25\}$ sites each. Estimate baseline linear model $\hat{\tau}(\mathbf{x}) = \mathbf{x}^T \hat{\boldsymbol{\beta}}$ for treatment effect prediction.

Treatment Effect Generation: For signal strength $\eta \in \{0.3, 0.66, 0.9\}$, generate individual effects:

$$\tau_i = \eta \cdot \text{standardize}(\hat{\tau}(\mathbf{x}_i)) + (1 - \eta) \cdot \varepsilon_i + \gamma U_i$$

Where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$, $U_i \sim \mathcal{N}(0, 1)$, and γ controls unmeasured heterogeneity. Population ATE: $\text{PATE} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \bar{\tau}_s$.

Distribution Shift: We shift site level covariates in the following way:

$$\mathbf{X}_s^{\text{shifted}} = \mathbf{X}_s + \rho \cdot \frac{d_{\text{med}}}{2} \cdot \frac{\mathbf{X}_s - \bar{\mathbf{X}}}{\|\mathbf{X}_s - \bar{\mathbf{X}}\|}$$

where

$$\rho \in \{0.0, 0.4, 0.6, 0.9, 1.7, 3.4\}, \quad d_{\text{med}} = \text{median}_{s,s'} \|\mathbf{X}_s - \mathbf{X}_{s'}\|$$

Site Selection Methods: From each N-site pool, select K sites using:

Random: Uniform sampling (averaged over 15 trials)

Stratification: k-means clustering + within-stratum sampling

SPS: Synthetic Purposive Sampling (Egami et al. 2024).

Wasserstein DRO: Variants combining PATE/CATE objectives ($p \in \{1, 2\}$) with robustness radius $\rho^* \in \{0, Q_{25}, Q_{50}, Q_{75}\}$ calibrated from empirical site distances

Robustness Calibration: Compute pairwise Wasserstein distances between sites. Set $\rho^* = 0$ (non-robust), 25th/50th/75th percentiles of observed distance in the data.

Performance Metrics:

PATE: $\text{MSE} = (\hat{\text{PATE}} - \text{PATE})^2$

CATE: $\text{PEHE} = \mathbb{E}[(\tau(\mathbf{x}) - \hat{\tau}(\mathbf{x}))^2]$ where $\hat{\tau}(\mathbf{x})$ is linear model fit on selected sites.

Output: Aggregate performance across $3 \times 6 = 18$ scenarios (signal × shift combinations), comparing method effectiveness under varying conditions.

C Implementation Details

C.1 LP Relaxations of the MILP and Cutting-Plane Algorithm

LP Relaxation of the MILP

In general, the LP relaxation of an MILP removes the ‘mixed integer’ constraint – instead of requiring that we solve an hard *discrete* optimization problem with binary indicators, we solve a relaxed version of the problem, where integers are allowed to take continuous values in $[0, 1]$, with rounding occuring after a solution to this problem has been found. Continuous linear programs can be solved in polynomial time, while integer programming is NP-hard (Karmarkar 1984; Natarajan 1995). The site inclusion indicators $s_i \in \{0, 1\}$ are relaxed to $s_i \in [0, 1]$.

LP Relaxation of the Cutting-Plane Algorithm

In the robust setting ($\rho > 0$), the cutting-plane algorithm alternates between adversarial distribution selection and site selection response. Now we solve two LPs in each iteration: the adversary maximizes transport cost subject to the Wasserstein budget constraint, then the decision maker minimizes maximum transport cost over all observed adversarial distributions.

Warm Starting

As a default, to speed up implementation, LP relaxation is used as initialization strategy for exact MILP solvers in both nonrobust and DRO settings. The continuous solution provides warm start values by initializing binary variables to rounded values of the relaxed solution, often reducing branch-and-bound iterations by orders of magnitude. For problems with $n > 100$ sites, LP relaxation is used as the default implementation, rather than as the warm start.

C.2 Runtime Experiments

Table 5: Runtime Comparison: Exact MILP vs LP Relaxation for 1-Transport

Sites	Selected	Combinations	Exact (s)	LP (s)	Speedup
10.00	3.00	1.200000e+02	0.295	0.064	4.6
15.00	4.00	1.365000e+03	0.143	0.077	1.9
20.00	5.00	1.550400e+04	0.304	0.119	2.6
25.00	6.00	1.771000e+05	0.316	0.127	2.5
30.00	7.00	2.0e+06	0.429	0.190	2.3
40.00	10.00	8.5e+08	1.416	0.391	3.6
50.00	12.00	1.2e+11	1.798	0.587	3.1
75.00	18.00	9.6e+16	5.742	1.953	2.9
100.00	25.00	2.4e+23	18.741	4.386	4.3
150.00	37.00	1.9e+35	616.924	16.755	36.8
200.00	50.00	4.5e+47	—	46.248	—

D Additional Theoretical Results

D.1 Optimal Transport and Survey Sampling

1-Wasserstein transport as balanced sampling on 1-Lipschitz functions

The 1-Wasserstein site selection problem is equivalent to balanced sampling that simultaneously controls the sampling error over the class of 1-Lipschitz functions. Intuitively, this tells us how we should think about the solution set: we choose the sites that are most likely to balance error over all 1-Lipschitz functions of the covariates.

Theorem 28 (1-Wasserstein Transport as Balanced Sampling)**.** *Let $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ be a finite population with uniform empirical measure $P_X = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. For any subset $S \subset \{1, \ldots, n\}$ with $|S| = K$, define $S_X = \frac{1}{K}\sum_{j \in S} \delta_{x_j}$.*

The 1-Wasserstein site selection problem

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

is equivalent to the balanced sampling problem

$$\min_{S\,|S|=K} \quad \sup_{f \in Lip_1(\mathbb{R}^d)} \left| \frac{1}{n}\sum_{i=1}^n f(x_i) - \frac{1}{K}\sum_{j \in S} f(x_j) \right|$$

where $Lip_1(\mathbb{R}^d) = \{f : \mathbb{R}^d \to \mathbb{R} : ||f||_{Lip} \leq 1\}$ is the class of 1-Lipschitz functions.

Proof. The equivalence follows directly from the Kantorovich-Rubinstein duality theorem for 1-Wasserstein distance.

By the Kantorovich-Rubinstein theorem, for any two probability measures μ, ν on a metric space $(\mathcal{X}, d)$:

$$W_1(\mu, \nu) = \sup_{f:||f||_{\text{Lip}} \leq 1} \left| \int f\, d\mu - \int f\, d\nu \right|$$

Applying this to our discrete measures P_X and S_X:

$$\begin{aligned} W_1(P_X, S_X) &= \sup_{f:||f||_{\text{Lip}} \leq 1} \left| \int f\, dP_X - \int f\, dS_X \right| \\ &= \sup_{f:||f||_{\text{Lip}} \leq 1} \left| \frac{1}{n}\sum_{i=1}^n f(x_i) - \frac{1}{K}\sum_{j \in S} f(x_j) \right| \end{aligned}$$

Therefore:

$$\min_S W_1(P_X, S_X) = \min_S \sup_{f:||f||_{\text{Lip}} \leq 1} \left| \frac{1}{n}\sum_{i=1}^n f(x_i) - \frac{1}{K}\sum_{j \in S} f(x_j) \right|$$

This establishes the claimed equivalence. □

Remark 29 (Comparison with Classical Balanced Sampling)**.** Classical balanced sampling typically balances on a finite set of auxiliary variables. The 1-Wasserstein formulation extends this to balance simultaneously over the infinite-dimensional class of all 1-Lipschitz functions.

2-Wasserstein transport as optimal stratified sampling

When the population size is divisible by the number of selected sites, 2-Wasserstein site selection is equivalent to optimal balanced stratified sampling. When the population size is not divisible by the number of selected sites, 2-Wasserstein site selection allows for fractional assignments, which strictly dominates optimal stratified sampling.

To prove this, I first show that the result holds in the case where N is divisible by K. I then show that an analogous optimality result holds when N is not divisible by K.

This equivalence helps us to understand why transport-based and stratification-based site selection methods for the CATE perform similarly in practice.

Theorem 30 (2-Wasserstein Transport as Optimal Stratification)**.** *Assume N is divisible by K. Let $\mathcal{X} = x_1, \ldots, x_n \subset \mathbb{R}^d$ be a finite population with uniform empirical measure $P_X = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$. For any subset $S \subset 1, \ldots, n$ with $|S| = K$, define $S_X = \frac{1}{K}\sum_{j \in S} \delta_{x_j}$. The 2-Wasserstein site selection problem*

$$\min_{S \subset 1,\ldots,n, |S|=K} W_2^2(P_X, S_X)$$

is equivalent to the optimal balanced stratification problem:

$$\min_{\mathcal{C},\mathbf{r}} \sum_{j=1}^{K} \sum_{i \in C_j} ||x_i - x_{r_j}||^2$$

where $\mathcal{C} = C_1, \ldots, C_K$ is a balanced partition of $1, \ldots, n$ with $|C_j| = \frac{n}{K}$ for all j, and $\mathbf{r} = (r_1, \ldots, r_K)$ with $r_j \in 1, \ldots, n$ for all j.

Proof. I establish equivalence by showing that optimal transport plans have a simple structure that corresponds exactly to balanced partitions.

The 2-Wasserstein problem requires solving:

$$\min_{\pi \in \Pi(P_X, S_X)} \sum_{i=1}^{n} \sum_{j \in S} \pi_{ij} ||x_i - x_j||^2$$

where $\Pi(P_X, S_X)$ contains transport plans satisfying marginal constraints.

Lemma 31 (Elements of optimal plan)**.** *Assume N is divisible by K. For any optimal transport plan π^*, we have $\pi^*_{ij} \in \{0, \frac{1}{n}\}$ for all (i, j).*

Proof. First, I show that the marginal constraints induce balanced partitions, then prove that plans with closest-site assignment dominate plans that assign mass fractionally.

Each population point i has mass $\frac{1}{n}$ and each selected site $j \in S$ must receive mass $\frac{1}{K}$. Since $\frac{1}{K} = \frac{N/K}{N}$, each selected site must receive mass from exactly $\frac{N}{K}$ population points.

Given the discrete uniform structure, any feasible transport plan must satisfy $\sum_{j \in S} \pi_{ij} = \frac{1}{N}$, for each site i – that is, that the mass of each site i must be fully allocated to sites j; and $\sum_{i=1}^{n} \pi_{ij} = \frac{1}{K}$ for each site j, that is, that each site j receives mass equal to $\frac{1}{K}$.

Since each population point has indivisible mass $\frac{1}{N}$ and each selected site requires mass from exactly $\frac{N}{K}$ points, any feasible transport plan corresponds to a partition of the population into K groups of size $\frac{N}{K}$.

Suppose for contradiction that some optimal plan π^* has $\pi^*_{ij} \in (0, \frac{1}{N})$ for population point i and selected sites $j, j' \in S$ with $j \neq j'$, so that point i fractionally splits its mass between j and j'.

However, any such fractional assignment can be improved by a reassignment that respects the marginal constraints. Since the marginal constraints require each population point to send its full mass $\frac{1}{N}$ somewhere, and splitting mass between distant points increases transport cost, the optimal strategy assigns each population point entirely to its closest selected site among those with remaining capacity. More precisely, any transport plan with fractional assignments can be converted to a partition-based plan with the same marginal totals but lower objective value by reassigning each population point entirely to its closest selected site, contradicting optimality. □

The optimal transport plan π^* induces a partition $C_j : j \in S$ where $C_j = i : \pi^*_{ij} = \frac{1}{N}$. The target marginal constraint ensures balance: $\sum_{i \in C_j} \frac{1}{N} = \frac{1}{K}$ implies $|C_j| = \frac{n}{s}$. The objectives are identical up to scaling:

$$W_2^2(P_X, S_X) = \frac{1}{n} \sum_{j \in S} \sum_{i \in C_j} ||x_i - x_j||^2$$

Now, I show that these problems are equivalent. Given optimal site selection S with transport plan π^*, construct stratification by setting $C_j = i : \pi^*_{ij} = \frac{1}{N}$ and $r_j = j$ for $j \in S^*$.

Conversely, given optimal stratification $(\mathcal{C}^*, \mathbf{r}^*)$, construct site selection $S^* = \{r_1^*, \ldots, r_s^*\}$ with transport plan $\pi^*_{ij} = \frac{1}{N}$ if $i \in C_k$ and $j = r_k$, zero otherwise. Both mappings preserve optimality and establish problem equivalence. □

Corollary 32 (Optimality versus Stratified Sampling)**.** *Assume N is divisible by K. 2-Wasserstein site selection weakly dominates any stratified sampling procedure that separates stratification and representative selection.*

Proof. Let $\mathcal{F}_{\text{standard}}$ denote the feasible set of standard stratification, which first fixes a partition $\mathcal{P}$ according to some criterion, then optimizes representatives within strata:

$$\mathcal{F}_{\text{standard}} = \{(\mathcal{P}, \mathbf{r}) : \mathcal{P} \text{ fixed by Stage } 1, r_j \in C_j \text{ for all } j\}$$

Let $\mathcal{F}_{\text{Wasserstein}}$ denote the feasible set of 2-Wasserstein optimization:

$$\mathcal{F}_{\text{Wasserstein}} = \{(\mathcal{P}, \mathbf{r}) : \mathcal{P} \text{ balanced partition}, r_j \in \{1, \ldots, n\} \text{ for all } j\}$$

Since stratification restricts representatives to lie within their assigned strata while 2-Wasserstein allows any population point as a representative, we have:

$$\mathcal{F}_{\text{standard}} \subset \mathcal{F}_{\text{Wasserstein}}$$

. Therefore:

$$\min_{(\mathcal{P},\mathbf{r}) \in \mathcal{F}_{\text{Wasserstein}}} \sum_{j=1}^{s} \sum_{i \in C_j} \|x_i - x_{r_j}\|^2 \leq \min_{(\mathcal{P},\mathbf{r}) \in \mathcal{F}_{\text{standard}}} \sum_{j=1}^{s} \sum_{i \in C_j} \|x_i - x_{r_j}\|^2$$

with equality when stratification produces the globally optimal solution. □

Remark 33. Standard stratification first fixes a partition, then optimizes representatives within strata. This restricts the feasible set compared to 2-Wasserstein optimization, which jointly optimizes partitions and representatives with the constraint that representatives come from the full population.

Remark 34 (Non-divisible case)**.** When $n \bmod K \neq 0$, the equivalence to balanced stratification no longer holds exactly. The optimal transport plan must use fractional assignments $\pi^*_{ij} \in [0, 1/n]$ to satisfy the marginal constraint $\sum_i \pi_{ij} = 1/K$ at each selected site (since n/K is non-integer). Meanwhile, stratified sampling is constrained to integer assignments with unequal stratum sizes. The dominance argument of Corollary 32 still holds: fractional assignments form a strictly larger feasible set than integer assignments, so the 2-Wasserstein solution achieves a lower objective value.

Remark 35 (CATE solution induces an Optimal Voronoi Partition of the Covariate Space)**.** The optimal solution creates constrained Voronoi cells where each cell contains exactly $\frac{n}{s}$ population points and centroids are chosen from the population to minimize total within-cell variance. We can interpret the Voronoi cells as optimal strata.

Remark 36 (Relationship to k-means clustering)**.** While k-means allows arbitrary centroids in $\mathbb{R}^d$, 2-Wasserstein transport constrains centroids to the original population and enforces balanced clusters when n is divisible by K. When $n \bmod K \neq 0$, it enforces clusters as balanced as the discrete constraint allows (sizes $\lfloor n/K \rfloor$ or $\lceil n/K \rceil$), making it a discrete, approximately-balanced variant of k-means clustering.

D.2 Game Theory and Distributionally Robust Optimization

We can interpret Distributionally Robust Optimization as a game played between Nature and a Researcher.

Setup

Consider the following game:

Actors

- A **Researcher**, who selects sites S to minimize representation error wrt P
- **Nature**, who perturbs the population distribution to maximize representation error

Order of Actions

1. The Researcher observes population sites $\{x_1, \ldots, x_n\}$ and chooses site selection $S \subseteq |P|$ with $|S| = K$
2. Nature observes the Researcher's choice and selects adversarial distribution Q subject to budget constraint $W_p(Q, P_X) \leq \rho$
3. Payoffs are realized based on representation error $W_p^p(Q, S_X)$

Action Spaces

$$\mathcal{A}_{\text{Researcher}} = \{S \subseteq [n] : |S| = s\}$$
$$\mathcal{A}_{\text{Nature}} = \{Q \in \mathcal{P}(\{x_1, \ldots, x_n\}) : W_p(Q, P_X) \leq \rho\}$$

Payoffs

The Researcher seeks to minimize representation error. Nature seeks to maximize it. The payoff function is:

$$u(S, Q) = W_p^p(Q, S_X)$$

where $S_X = \frac{1}{s} \sum_{j \in S} \delta_{x_j}$ is the empirical distribution of selected sites.

The Researcher receives payoff $-u(S, Q)$ and Nature receives payoff $u(S, Q)$ (this is a zero-sum game).

Equilibrium Analysis

Definition 37 (Subgame Perfect Equilibrium)**.** The subgame perfect equilibrium $(S^*, Q^*(\cdot))$ satisfies:

Nature's Best Response: For any $S \in \mathcal{A}_{\text{Researcher}}$,

$$Q^*(S) \in \arg\max_{Q \in \mathcal{P}(\{x_1,\ldots,x_n\})} \left\{ W_p^p(Q, S_X) : W_p(Q, P_X) \leq \rho \right\}$$

Researcher's Optimal Strategy:

$$S^* \in \arg\min_{S \in \mathcal{A}_{\text{Researcher}}} W_p^p(Q^*(S), S_X)$$

The equilibrium value is:

$$V^* = \min_{S \subseteq [n], |S| = s} \max_{Q : W_p(Q, P_X) \leq \rho} W_p^p(Q, S_X)$$

Variable Interpretation

Variable	**Interpretation**
$z_j \in \{0,1\}$	Site selection indicator
$\mu_k \geq 0$	Nature's adversarial distribution
$\alpha_{ik} \geq 0$	Transport from original to adversarial distribution
$\beta_{kj} \geq 0$	Transport from adversarial to selected distribution

Mixed-Integer Linear Program Formulation

The equilibrium can be computed by solving:

$$\min_{z,\mu,\alpha,\beta} \sum_{k=1}^{n} \sum_{j=1}^{n} \beta_{kj} d(x_k, x_j)^p \tag{1}$$

$$\text{subject to} \quad \sum_{j=1}^{n} z_j = s \tag{2}$$

$$\sum_{k=1}^{n} \mu_k = 1 \tag{3}$$

$$\sum_{k=1}^{n} \alpha_{ik} = \frac{1}{n} \quad \forall i \tag{4}$$

$$\sum_{i=1}^{n} \alpha_{ik} = \mu_k \quad \forall k \tag{5}$$

$$\sum_{j=1}^{n} \beta_{kj} = \mu_k \quad \forall k \tag{6}$$

$$\sum_{k=1}^{n} \beta_{kj} = \frac{z_j}{s} \quad \forall j \tag{7}$$

$$\beta_{kj} \leq z_j \quad \forall k, j \tag{8}$$

$$\sum_{i,k} \alpha_{ik} d(x_i, x_k)^p \leq \rho^p \tag{9}$$

$$z_j \in \{0,1\}, \quad \mu_k, \alpha_{ik}, \beta_{kj} \geq 0 \tag{10}$$

Constraints

Linking Constraint (8): If site j is not selected ($z_j = 0$), then $\beta_{kj} = 0$ for all k. Nature cannot assign transport cost to unselected sites.
Researcher's Budget Constraint (2): Researcher can choose K sites.
Nature's Budget Constraint (9): Limits Nature's ability to perturb the distribution. Larger ρ gives Nature more power to create challenging distributions.
Transport Constraints (4)-(7): Ensure valid probability distributions and transport plans.

Discussion

This game theoretic formulation motivates the cutting-plane algorithm described in Section 3.3: the Researcher chooses sites, Nature responds with worst-case distribution, the Researcher updates their site selection based on all perturbations observed so far, and the process continues until convergence to Nash equilibrium. This is an illustration of an algorithm that implements fictitious play (Brown 1951; Roughgarden 2016).