WHOMP: Optimizing Randomized Controlled Trials via Wasserstein Homogeneity

Shizhou Xu SHZXU@UCDAVIS.EDU
Department of Mathematics
University of California Davis
Davis, CA 95616-5270, USA

Thomas Strohmer STROHMER@MATH.UCDAVIS.EDU
Department of Mathematics
Center of Data Science and Artificial Intelligence Research
University of California Davis
Davis, CA 95616-5270, USA

Abstract

We investigate methods for partitioning datasets into subgroups that maximize diversity within each subgroup while minimizing dissimilarity across subgroups. We introduce a novel partitioning method called the *Wasserstein Homogeneity Partition* (WHOMP), which optimally minimizes type I and type II errors that often result from imbalanced group splitting or partitioning, commonly referred to as accidental bias, in comparative and controlled trials. We conduct an analytical comparison of WHOMP against existing partitioning methods, such as random subsampling, covariate-adaptive randomization, rerandomization, and anti-clustering, demonstrating its advantages. Moreover, we characterize the optimal solutions to the WHOMP problem and reveal an inherent trade-off between the stability of subgroup means and variances among these solutions. Based on our theoretical insights, we design algorithms that not only obtain these optimal solutions but also equip practitioners with tools to select the desired trade-off. Finally, we validate the effectiveness of WHOMP through numerical experiments, highlighting its superiority over traditional methods.

Keywords: randomized controlled trial, Wasserstein homogeneity, anti-clustering, diverse K-means, control/test group splitting, cross-validation

Contents

1. Introduction

Congratulations! After investing years of hard work and hundreds of millions of dollars, your company has discovered a promising new cancer drug. The next milestone is to conduct a *randomized clinical trial* to confirm the drug's effectiveness. However, occasionally the randomization procedure can cause an imbalance in covariates related to the outcome across groups. A chance you are reluctant to take, since too much is at stake here! You are, of course, aware of the various attempts to mitigate the potential downsides of randomization, such as covariate adaptive randomization. But these alternatives have their own drawbacks, often seem ad hoc, and very rare of these methods are designed with any optimality criteria for comparative tests. Enter WHOMP, *Wasserstein HOMogeneous Partitioning*, a method that constructs maximally balanced data partitions with provable optimality guarantees.

1.1 Motivation

Randomized group splitting has been a widely accepted standard for estimating causal inference in scientific experiments, as randomization typically balances covariate effects between group divisions and experimental outcomes on average over repeated trials. However, the risks associated with pure randomization and imbalanced group splitting have been highlighted in numerous studies across fields such as agriculture, biology, social sciences, and clinical research [18, 23, 44, 34].

The widely held belief behind randomization is that it promotes comparability between the resulting subgroups. For instance, Rosenberger states in [33], "The first property of randomization is that it promotes comparability among the study groups." However, this result holds with reasonably high probability only when the law of large numbers applies to the randomized subsampling process. In many controlled trials, the sample size is inherently limited. Additionally, conducting repeated experiments with randomized sample splitting can be prohibitively expensive or even impractical in many scientific settings. Therefore, the law of large numbers may not apply to either the group size or the number of trials. As stated by Fisher, who first proposed the requirement of randomization in experimental design, in [18], "Most experimenters on carrying out a random assignment of plots will be shocked to find how far from equally the plots distribute themselves". When the (sub)sample size is insufficient or the number of covariates is relatively large, there is a non-negligible chance that the randomization itself becomes a covariate factor in a limited number of realizations, potentially leading to type I or type II errors.

This work aims to address the following question, which naturally arises in scientific experiments and causal inference studies:

> *How can we split a sample into control and test (or multiple controlled) groups in a way that minimizes the impact of the data splitting on the outcomes of the controlled experiment?*

We approach this question from two key perspectives:

- *In-subgroup diversity*: Maximizing diversity within each subgroup (or partition element) based on a specific diversity metric.

- *Cross-subgroup similarity*: Minimizing dissimilarity across subgroups using a defined similarity measure.

Here, maximizing diversity within each subgroup ensures that the test results are more representative of the entire sample, which is often assumed to reflect the target population. At the same time, minimizing dissimilarity between subgroups, where different controlled factors are applied, reduces the likelihood that the statistical (in)significance is driven by covariate imbalances introduced through group splitting.

Beyond scientific experiments, the study of group splitting to maximize in-subgroup diversity and cross-subgroup similarity has garnered attention in various fields: *Graph Theory:* Partitioning the nodes of a (weighted) graph into clusters such that the total weight of edges with both endpoints in the same cluster is maximized [17, 16]. *Federated Learning:* Identifying "superclients" to address distribution heterogeneity in training data across clients, using either unsupervised approaches [26, 45] or supervised methods [12]. *Managerial Science:* Promoting diversity within workgroups to enhance productivity [4, 6, 15].

In this work, we propose a new partitioning objective that addresses both perspectives:

> *Homogeneity Partition*: Given a distance metric on probability distribution spaces, such as Wasserstein spaces, the partitioning method aims to minimize the average squared distance between the entire sample and the resulting subgroups.

Here, in-subgroup diversity is captured by minimizing the distance between the subgroup and the entire sample: less diversity (relative to the entire sample) in a subgroup results in a greater distance between it and the entire sample. On the other hand, cross-subgroup similarity is captured by the minimization of average squared distance: The average squared distance minimizes the variability among the subgroups around the sample. The distance is squared because ℓ^2-minimization promotes a more uniform distribution of distance or variability, compared to ℓ^1-minimization, and a more balanced distribution of the distributional metric results in cross-subgroup similarity.

In this study, we concentrate on the Wasserstein-2 distance and present a comprehensive analysis of the above considerations. The main contributions of our work are as follows:

- **Optimality criterion** (Section 2): We introduce a novel partitioning objective, named WHOMP, which is designed to establish a provable optimality criterion that guarantees the effectiveness of comparative experiments through the resulting partitions.

- **Analytical comparison** (Section 3): We provide a thorough analytical comparison of widely used partitioning methods in comparative experiments, including random partitioning, covariate-adaptive randomization, rerandomization, and anti-clustering, with the proposed WHOMP. This analysis highlights the connections, distinctions, and advantages of WHOMP in comparison to existing methods.

- **Solution characterization and algorithm design** (Sections 4, 5): We characterize the optimal solutions to the WHOMP problem and develop an efficient algorithm for their estimation. Furthermore, we identify a trade-off among different WHOMP optimal solutions, offering guidance for practitioners in selecting the most appropriate solution for specific use cases.

- **Numerical Comparison** (Section 6): We perform a numerical comparison of the standard partitioning methods and the WHOMP (implemented with our algorithm design) across various data types, including tabular, image, and graph data.

1.2 Related Work

One line of research related to diversified subgroup generation involves balancing significant covariates during randomized group splitting. This idea goes back at least to [14] and has been widely employed in various comparative studies, including clinical trials [33], A/B testing for business decisions [42], and experiments in the social sciences [13]. The objective is to balance covariates that may influence the results during randomized group splitting, thereby enhancing the credibility and efficiency of the trial or experiment. In other words, the goal is to reduce type I and type II errors caused by covariate imbalances in group assignments. To achieve this, methods such as covariate-adaptive randomization [25], block-stratified randomization [22], and minimization [35, 9] are commonly employed. Despite their extensive application, these methods have faced criticism for lacking optimality criteria related to guaranteed comparative test performance [36, 34, 21]. Existing methods either reduce distributional similarity to similarity in the first moments, as in minimization [39, 31], or rely on the assumption of a specific model for treatment effects [3].

Another line of research related to this work focuses on maximizing in-subgroup diversity, though these problems are studied under various terms, such as *anti-clustering* [37, 30], *K-partition* [16], *equitable partition* [29], and *maximally diverse group problem* [8, 19, 32]. The distinction among these problems is that some consider a more general distance or diversity penalty function beyond the Euclidean distance or variance. It is important to note that when enforcing uniform cardinality of subgroups, all these problems are equivalent in the Euclidean setting. Therefore, we use the term anti-clustering to represent this body of work and explore the similarities and differences between WHOMP and anti-clustering to highlight the advantages of the proposed method.

In particular, we highlight a common misunderstanding in the current anti-clustering approach to the diverse subgroups problem. As discussed earlier, the problem of subgroup splitting for comparative tests should encompass two aspects that are not necessarily compatible: *In-subgroup diversity* and *cross-subgroup similarity*. The anti-clustering approach primarily focuses on in-subgroup diversity but relies on the following duality result to argue that maximizing in-subgroup diversity also maximizes cross-subgroup similarity: Maximizing in-subgroup variance is equivalent to minimizing the variance of the centroids across subgroups. For the exact statement, see Lemma 3.4 or [37]. That is, this equivalence holds only when cross-subgroup similarity is defined as similarity in subgroup averages.

However, enforcing similarity among subgroup averages often leads to scale differences among subgroups. Points with similar scales but different directions tend to be grouped together to balance each other out, thereby achieving similarity in averages. While this scale matching can be beneficial for certain applications, the resulting cross-group scale differences make the anti-clustering objective less suitable when distributional properties or higher-order statistics are of greater concern than simple expectations. We demonstrate that the proposed objective in this work is more suitable for scenarios where similarity in distribution or higher-order statistics (beyond simple expectations) is of greater importance.

Rerandomization [38, 27] is another line of work that aims to address imbalanced covariate issues in partitioning control and test subgroups while maintaining the robustness that stems from randomness. In a standard rerandomization approach, a quantification for covariate imbalance is first established along with a threshold for accepting or rejecting subgroup splits. A random partition is then generated, and the covariate imbalance is computed. Based on whether the imbalance falls below the threshold, the partition is either accepted or rejected. The typical imbalance measures used in rerandomization include average difference or Mahalanobis distance, which primarily focus on the average similarity between subgroups but overlook other important distributional discrepancies. Moreover, the threshold for accepting or rejecting a partition is often determined manually, which introduces subjectivity into the process. In contrast, as shown in Section 3, WHOMP can be implemented as a rerandomization strategy that overcomes these limitations. Specifically, it utilizes optimal transport to design an imbalance metric that captures broader distributional discrepancies between subgroups, and it employs unsupervised learning techniques to automatically determine the threshold based on the covariates.

A different, very interesting, approach is taken in [20]. There, the authors first formalize the tradeoff between covariate balance and a notion of robustness. By linking the experimental design problem to a new type of problem in algorithmic discrepancy, the authors then propose a randomized algorithm, namely the Gram-Schmidt walk, to solve the distributional discrepancy problem and thereby navigate the tradeoff between balance and robustness. Their method is limited to the specific setting, where one aims to split the data set into *two* subgroups.

Finally, this work is also related to statistical parity in machine learning group fairness. Since a covariate-balanced partition can be viewed as one that is independent of the covariate, its objective aligns with the definition of statistical parity in ML group fairness. In fact, the proposed solution for WHOMP in this work is closely related to performing K-means clustering on an optimal fair data representation that ensures statistical parity. This fair data representation approach guarantees statistical parity for any neutral downstream tasks, such as K-means clustering [43]. Here, neutral downstream tasks are the downstream tasks or models that do not introduce statistical dependence on the sensitive information by themselves.

1.3 Preliminaries and Notation

We provide a brief review of optimal transport, the Wasserstein-2 space[1], and the Wasserstein barycenter, which are essential tools in the development and analysis of WHOMP.

Given $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ where $\mathcal{P}(\mathbb{R}^d)$ denotes the set of all the probability measures on $\mathbb{R}^d$,

$$\mathcal{W}_2(\mu,\nu) := \left(\inf_{\lambda \in \prod(\mu,\nu)} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} ||x_1 - x_2||^2 d\lambda(x_1, x_2) \right\} \right)^{\frac{1}{2}}.$$

Here, $\prod(\mu,\nu) := \{\pi \in \mathcal{P}((\mathbb{R}^d)^2) : \int_{\mathbb{R}^d} d\pi(\cdot, v) = \mu, \int_{\mathbb{R}^d} d\pi(u, \cdot) = \nu\}$. $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is called the Wasserstein space, where $\mathcal{P}_2(\mathbb{R}^d) := \left\{ \mu \in \mathcal{P}(\mathbb{R}^d) : \int_{\mathbb{R}^d} ||x||^2 d\mu < \infty \right\}$. To simplify

1. As this work focuses on the Wasserstein-2 space, we will henceforth refer to it simply as the Wasserstein space

notation, we often denote

$$\mathcal{W}_2(X_1, X_2) := \mathcal{W}_2(\mathcal{L}(X_1), \mathcal{L}(X_2)),$$

where $\mathcal{L}(X) := \mathbb{P} \circ X^{-1} \in \mathcal{P}(\mathbb{R}^d)$ is the law or distribution of X, $X : \Omega \to \mathcal{X} := \mathbb{R}^d$ is a random variable (or vector) with an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Intuitively, one can consider the Wasserstein distance as L^2 distance after optimally coupling two random variables whose distributions are μ and ν. That is, if the pair (X_1, X_2) is an optimal coupling [40], then

$$\mathcal{W}_2(X_1, X_2) = ||X_1 - X_2||_{L^2} = \int_\Omega ||X_1(\omega) - X_2(\omega)||^2 d\mathbb{P}(\omega).$$

Given $\{\mu_z\}_{z\in\mathcal{Z}} \subset (\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ for some index set $\mathcal{Z}$, their Wasserstein barycenter [1] with weights $\lambda \in \mathcal{P}(\mathcal{Z})$ is

$$\bar{\mu} := \text{argmin}_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \Big\{ \int_{\mathcal{Z}} \mathcal{W}_2^2(\mu_z, \mu) d\lambda(z) \Big\}. \tag{1}$$

If there is no danger of confusion, we will refer to the Wasserstein barycenter simply as barycenter.

Two random variables X_1 and X_2 are called equal in distribution if they have the same probability distribution, which is denoted by $X_1 =_d X_2$. More specifically, $X_1 =_d X_2$ if and only if, for all $f \in \mathcal{C}_b(\mathbb{R}^d)$,

$$\int_{\mathbb{R}^d} f d\mathcal{L}(X_1) := \int_{\mathbb{R}^d} f(x) d\mathbb{P} \circ (X_1)^{-1}(x) = \int_{\mathbb{R}^d} f(x) d\mathbb{P} \circ (X_2)^{-1}(x) =: \int_{\mathbb{R}^d} f d\mathcal{L}(X_2),$$

where $\mathcal{C}_b(\mathbb{R}^d)$ denotes the set of all bounded continuous functions on $\mathbb{R}^d$.

The rest of this paper is organized as follows: Section 2 defines the Wasserstein Homogeneity Partition (WHOMP) and shows that the WHOMP objective is desirable in group splitting for comparative experiments, such as clinical trials, business A/B tests, and social studies. Section 3 provides a detailed comparison between WHOMP and other partition methods: random partition, stratified randomization, covariate-adaptive randomization, rerandomization, and anti-clustering, which shows the advantage of WHOMP. Section 4 proves a characterization of the WHOMP solution. Section 5 provides an efficient method to estimate the solution to WHOMP which leads to the design of a practical algorithm. Finally, Section 6 demonstrates the advantages of WHOMP via numerical experiments on various data sets.

2. WHOMP: Wasserstein Homogeneity Partition

In this section, we propose the homogeneity partition: given a metric on the probability space, the homogeneity partition aims to minimize the sum of the squared distances between the resulting subgroups and the original data set. Furthermore, we show that, when applying the Wasserstein-2 distance to the homogeneity partition, we can provably minimize the Type I and II error due to the covariate factors resulting from the subgroup partition.

2.1 Definition of Wasserstein homogeneity

To start, we define the Wasserstein homogeneity partition for a given data set $X := \{x_i\}_{i\in[N]} \in \mathcal{X}^{[N]}$. To fix the notation below, we let $\mathbf{P}(N,K)$ denote all the partitions on $[N]$ that have K non-empty elements. That is,

$$P \in \mathbf{P}(N,K) \implies P = \{p_i\}_{i\in[K]} \text{ such that } \bigcup_{i\in[K]} p_i = [N] \text{ and } p_i \cap p_j = \emptyset, \forall i \neq j. \quad (2)$$

Also, given a partition $P = \{p_i\}_{i\in[K]}$ on X, we define $X_p := \{x_i\}_{i\in p}$ for all $p = p_i \in P$ and $X_P := \{X_p\}_{p\in P} = \{X_{p_i}\}_{i\in[K]}$, which is a set of the X_p's indexed by $[K]$. We also use p_i's to denote the following indicator functions:

$$p_i(j) = \begin{cases} 1 & \text{if } j \in p_i \\ 0 & \text{if } j \notin p_i \end{cases}$$

Similarly, we often use $P : [N] \to [K]$ to denote the map that $P(i) = j$ if and only if $x_i \in p_j$. Now, we are ready to define the Wasserstein homogeneity partition:

Definition 2.1 (Wasserstein Homogeneity Partition) *Given a data set $X := \frac{1}{N}\sum \delta_{x_i} \in \mathcal{P}(\mathcal{X})$ where $N = K \cdot c$ with $K, c \in \mathbb{N}$, the Wasserstein homogeneity partition problem is defined as*

$$\min_{\substack{Q\in\mathbf{P}(N,c)\\ |q|\equiv K}} \sum_{q\in Q} \mathcal{W}_2^2(X_q, X). \quad (3)$$

Here, since the goal of the partition is to have data splitting for controlled experiments or A/B tests, we require uniform cardinality across the resulting partition subgroups in order to prevent sample ratio mismatch which is a common cause of Simpson's paradox. Moreover, when N does not divide K and c, the uniform cardinality constraint can be adjusted in practice or algorithm design to accommodate minimum and maximum cardinality requirements.

One can replace the $\mathcal{W}_2$-distance with some other notion of distance between X_q and X that may be better suited for particular applications. In this paper, we will focus on $\mathcal{W}_2$ because it is closely related to variance and L^2-loss, which leads to straight-forward solutions to WHOMP via K-means clustering.

We will demonstrate in Section 4 that finding the Wasserstein Homogeneity Partition, i.e., computing the solution of (3) is actually NP-hard. Yet, the reader need not despair, since our approach to showing NP-hardness also points to a convenient way to sidestep the NP-hardness. Indeed, we will see in Theorem 4.1 that solving (3) is equivalent to finding the solution to the (balanced) K-means clustering problem:

Definition 2.2 (Balanced K-means Clustering) *Given the data set $X := \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i} \in \mathcal{P}(\mathcal{X})$, the Balanced K-means Partition problem is defined as*

$$P \in \operatorname*{arg\,min}_{\substack{P\in\mathbf{P}(N,K),\\ |p|\equiv c}} \sum_{p\in P} \mathrm{Var}(X_p), \quad (4)$$

Based on this connection we will later present an approximate solution to Problem (3) by combining a balanced K-means approximation algorithm and randomness.

2.2 Theoretical guarantees for WHOMP

Here, we show that the objective function of the proposed homogeneity partition makes it suitable for comparative experiments such as clinical trials, A/B tests, and social science studies: The objective tends to penalize the average distributional discrepancy (quantified by the Wasserstein distance) between the subgroups and the original sample data and, hence, minimize the influence of the subsampling process on the controlled experiments outcome.

In particular, the following results provide a comprehensive justification for the WHOMP objective as a natural subsampling approach for statistical tests in comparative experiments, presented from three perspectives: qualitative (Theorem 2.1), concrete (Example 3.1 and the results therein), and quantitative (Corollary 2.2 and Theorem 2.2). To fix ideas for the following results, let X be the available feature variables in the sample, C the control variable, Q the subsampling assignment variable, and Y the controlled experiment outcome. To simplify notation, we assume the experiment will apply the control factor c_i to the subgroup X_{q_i} for all $i \in [c]$.

To start, we demonstrate that a zero WHOMP objective eliminates Type I and Type II errors arising from the subgroup splitting variable Q in the context of statistical or social experiments, which motivates the use of the Wasserstein homogeneity criterion:

Theorem 2.1 (No type I or type II error due to subgroup) *Assume that, for all $q \in Q$, $\mathcal{W}_2^2(X_q, X) = 0$. Also, let $Y : \mathcal{X} \times \mathcal{C} \to \mathcal{X}$ be the true outcome, which we assume to be an arbitrary measurable function. Then the following holds:*

- *For all $c_0, c_1 \in \mathcal{C}$ satisfying $Y(X, c_1) =_d Y(X, c_0)$, $Y(X_{q_1}, c_1) =_d Y(X_{q_0}, c_0)$*
- *For all $c_0, c_1 \in \mathcal{C}$ satisfying $Y(X, c_1) \neq_d Y(X, c_0)$, $Y(X_{q_1}, c_1) \neq_d Y(X_{q_0}, c_0)$*

Proof See Appendix A. ■

The above result shows that, for any ineffective treatment on the population X, the control/test experiment under (X_{q_0}, X_{q_1})-splitting reveals the ineffectiveness truthfully. In other words, no type I error arises from subgroup partitioning or the subgroup variable Q. Similarly, for any treatment that is genuinely effective for the population, the control/test experiment under (X_{q_0}, X_{q_1}) partitioning must demonstrate this effectiveness. Specifically, there exists a test function $f \in \mathcal{C}_b(\mathcal{X})$ such that, when conducting hypothesis testing on the average effect of f, no type II error is introduced by the subgroup variable due to group-splitting. Although, for ease of presentation, Theorem 2.1 is stated in the case where the subgroups and controls are binary, it is clear that our problem setting is suitable for arbitrary discrete or continuous subgroups or control variables.

Next, we demonstrate the effect of zero WHOMP objective in hypothesis testing for comparative experiments via the following example:

Example 2.1 (Theorem 2.1 in Hypothesis Testing) Here, we use a standard hypothesis testing setting to illustrate Theorem 2.1 above and motivate the Theorem 2.2 below,

under the assumption that the experimenter is interested in estimating the average treatment effect:

$$\tau := \frac{1}{n}\sum_{i=1}^{n} y_i(0) - \frac{1}{n}\sum_{i=1}^{n} y_i(1), \tag{5}$$

where $y_i(j) := Y(x_i, c_j)$ for $j \in \{0, 1\}$. In experiments, it is not possible to observe the effects of different controlled factors on the same x_i. Therefore, we employ the following natural estimator to approximate τ:

$$\hat{\tau} := \bar{Y}_{\text{obs}}(0) - \bar{Y}_{\text{obs}}(1), \tag{6}$$

where $\bar{Y}_{\text{obs}}(0) := \frac{\sum_i y_i'(0)\mathcal{Q}(i)}{\sum_i \mathcal{Q}(i)}$ and $\bar{Y}_{\text{obs}}(1) := \frac{\sum_i y_i'(1)[1-\mathcal{Q}(i)]}{\sum_i [1-\mathcal{Q}(i)]}$. Here, $\mathcal{Q} : [N] \to \{0, 1\}$ represents the randomized partitions drawn from $\mathcal{Q}(P)$ (Definition 4.1) or, equivalently, resulting from WHOMP Random (see definition in Algorithm 1, Section 5).

The following result demonstrates that WHOMP Random produces an unbiased estimator:

Lemma 2.1 ($\hat{\tau}$ is an unbiased estimator of τ)

$$\mathbb{E}(\hat{\tau}) = \tau. \tag{7}$$

See proof in Appendix A. Now, we proceed with the standard steps for hypothesis testing:

(i) *Null Hypothesis*: Assume the null hypothesis of no average treatment effect, expressed as $\tau = 0$ because $y_i(0) := Y(x_i, c_0) = Y(x_i, c_1) =: y_i(1), \forall i \in [n]$.

(ii) *Null Distribution*: Generate the empirical estimator or null distribution, defined by the law of $\hat{\tau}(\mathcal{Q})$: $\mathcal{L}(\hat{\tau}(\mathcal{Q})|\mathcal{Q}(P)) := \{\frac{\sum_i y_i(0)\mathcal{Q}(i)}{\sum_i \mathcal{Q}(i)} - \frac{\sum_i y_i(1)[1-\mathcal{Q}(i)]}{\sum_i [1-\mathcal{Q}(i)]}, \mathcal{Q} \sim \mathcal{Q}(P)\}$.

(iii) *p-Value*: Compute the p-value as the frequency of equally or larger absolute value in $\mathcal{L}(\hat{\tau}(\mathcal{Q}))$ compared to the absolute value of the actual experimental observation.

The following result demonstrates how Theorem 2.1 applies within this hypothesis testing framework:

Corollary 2.1 (Zero-One p-value) *If $\mathcal{W}_2^2(X_i, X) = 0$ for $i \in \{0, 1\}$, it follows that*

$$\mathcal{L}(\hat{\tau}) = \delta_\tau = \delta_0. \tag{8}$$

In other words, the p-value is either 0 *or* 1.

Proof Since X and Q are finite, let $L := \max_{x,x' \in \mathcal{X}} \max_{q,q' \in Q} \min\{L : ||Y(x, c_q) - Y(x', c_{q'})||_{l^2} \leq L||x - x'||_{l^2}\}$. Notice the L is well-defined due to the null hypothesis that $Y(x, c_q) = Y(x, c_{q'}), \forall q, q' \in Q, \forall x \in X$. In addition, it follows from the null hypothesis and Lemma 2.1 that $\tau = 0$. Finally, it follows from Theorem 2.2 that, for all $\epsilon > 0$, we have

$$\text{Var}(\hat{\tau}) = \mathbb{E}(||\hat{\tau} - \tau||_{l^2}^2) \leq L^2 \frac{|Q|}{|Q| - 1} \sum_{q \in Q} \mathcal{W}_2^2(X_q, X) < \epsilon. \tag{9}$$

Therefore, we have $\mathrm{Var}(\hat{\tau}) = 0$, which implies $\mathcal{L}(\hat{\tau}) = \delta_0$. ■

Intuitively, the WHOMP objective is an effective tool for controlling the concentration (or standard deviation) of the null distribution around the true average treatment effect, especially when the observations are statistically dependent on the given covariates. Such control of concentration is crucial in scenarios where experiments can only afford to repeat the random trial a limited number of times. In the case described, measurability (with respect to (X, C)) and a zero WHOMP objective lead to the null distribution collapsing into a Dirac measure. See Theorem 2.2 for a more detailed result on the control of the distributional concentration of $\hat{\tau}$, under the more general assumptions that Y is Lipshcitz with respect to X.

One can also perform other hypothesis tests using the Wasserstein distance instead of the average distance. Specifically, one can use $\mathcal{W}_2^2(Y(X_0, c_0), Y(X_1, c_1))$ as an estimator for $\mathcal{W}_2^2(Y(X, c_0), Y(X, c_1))$. It can be shown that

$$\mathbb{E}(\hat{\tau}) = \mathbb{E}(\mathcal{W}_2^2(Y(X_0, c_0), Y(X_1, c_1))) = 4\mathbb{E}(\mathrm{Var}(X_P)) \neq 0 = \tau. \tag{10}$$

Therefore, $\mathbb{E}(\mathrm{Var}(X_P))$, which is proportional to the WHOMP objective, determines the bias in this case. For further details on hypothesis testing based directly on Wasserstein variation, we refer interested readers to [28, 10, 11] and the references therein.

Next, we demonstrate how the WHOMP objective bound the statistics estimation error between the resulting subgroups and the original sample:

Corollary 2.2 (Lipschitz Statistics Error Bound) *Assume* $\frac{1}{|Q|}\sum_{q\in Q}\mathcal{W}_2^2(X, X_q) \leq d$ *for some* $d \geq 0$*, then for any* $\epsilon > 0$*, and* $h : \mathcal{X} \to \mathbb{R}$*, we have*

$$\mathbb{P}(\{\sup_{||h||_{Lip}\leq L} |\mathbb{E}(h(X)) - \mathbb{E}(h(X_q))| > \epsilon\}) \leq \frac{L\sqrt{d}}{\epsilon}.$$

Proof See Appendix A. ■

Here, $||h||_{Lip} \leq L$ means that h is L-Lipschitz for some $L \in \mathbb{R}_+$. The above result shows that if the objective of Problem (3) (averaged by $|Q|$) is bounded by some d that is relatively small compared to $\frac{\epsilon}{L}$ for the chosen ϵ, then it is unlikely to observe that any L-Lipschitz statistics on X and X_q differ by more than ϵ.

Finally, we provide a quantitative version of Corollary 2.1: How the WHOMP objective controls the concentration of the average treatment effect estimator (unbiased by Lemma 2.1) around the true average effect?

Theorem 2.2 (Variance Bound for Average Treatment Effect Estimator) *Assume the observation is a uniformly (with respect to the control factor* $\{c_q\}_{q\in Q}$*) Lipschitz function of the given covariate* X*:*

$$\sup_{q,q'\in Q} ||Y(x, c_q) - Y(x', c_{q'})||_{l^2} \leq L||x - x'||_{l^2}, \forall x, x' \in \mathcal{X}, \tag{11}$$

then we have

$$\mathbb{E}(||\hat{\tau} - \tau||_{l^2}^2) \leq L^2 \frac{|Q|}{|Q| - 1} \sum_{q \in Q} \mathcal{W}_2^2(X_q, X), \forall Q \in \mathcal{Q}(P). \tag{12}$$

Proof See Appendix A ■

Without further assumptions on X, the inequality above is sharp as equality can be achieved by Gaussian mixture models.

3. Comparison with Related Subsampling Partition Methods

In this section, we provide an analysis of random subsampling, covariate-adaptive randomization, rerandomization, and anti-clustering, compared with WHOMP.

3.1 Random Sampling

Now, we show that pure random subsampling can result in large distributional deviations from the original sample, especially in the case of small subsample sizes. It is easy to construct examples with specific assumptions on sample distribution. For example, given a linear model $Y = ax + N$ where $N \sim \mathcal{N}(b, \sigma^2)$ is the Gaussian noise, one can consider subsamples as i.i.d. random variables drawing from Y and thereby conclude it is not unlikely to observe subsamples that significantly differ from Y.

Here, we give a result on the subsample deviation in terms of Wasserstein distance without assuming the sample distribution. Instead, we assume a quantity that we will need in our main result to characterize the solution to problem 3.1. In particular, we first show a deterministic lower bound when the subsample size is small, which is sharp and closely related to the Theorem 4.1 below, and thereby show that it is not unlikely to obtain large distributional deviations.

Lemma 3.1 (Subsample Wasserstein Deviation Lower Bound) *Let $X = \{x_i\}_{i=1}^N$ be a sample data set and $X_{sample} := \{x_i\}_{i=1}^K$ a subsample, where $x_i \sim uniform(X)$ are i.i.d. random variables sampled from X with uniform distribution. Then*

$$\mathcal{W}_2^2(X, X_{sample}) \geq \min_{\substack{P \in \mathbf{P}(N,K) \\ |p| \equiv c}} \frac{1}{K} \sum_{p \in P} \mathrm{Var}(X_p). \tag{13}$$

Proof See Appendix B. ■

Without further assumptions on the distribution of X, the lower bound is sharp due to the optimal partition P definition. But now we show that this lower bound is very unlikely to be obtained via random partition in practice and one should expect a much higher Wasserstein deviation with high probability:

Corollary 3.1 (Distribution of Subsample Wasserstein Deviation Lower Bounds) *With probability* $1 - \frac{K!}{K^K}$*, we have*

$$\mathcal{W}_2^2(X, X_{sample}) \geq \min_{\substack{P \in \mathbf{P}(N,K) \\ |p| \equiv c}} \frac{1}{K} \sum_{p \in P} \mathrm{Var}(X_p) + \min_{p \neq p'} ||\bar{X}_p - \bar{X}_{p'}||_2^2. \tag{14}$$

The distribution of the lower bounds is equal to drawing $\{I_k\}_{k=1}^K$ i.i.d. from $[K]$ uniformly with replacement and then sum $\min_{\substack{P \in \mathbf{P}(N,K), \frac{1}{K}\sum_{p\in P}\mathrm{Var}(X_p) \\ |p| \equiv c}}$ and

$$\min_{\sigma} \sum_{\substack{i: I_k \neq i, \forall k \\ j: |\{k: I_k = j\}| > 1}} ||\bar{X}_{p_i} - \bar{X}_{p_{\sigma(j)}}||_2^2. \tag{15}$$

Last but not least, it is important to note that the strength of randomized subsampling lies in its ability to enhance robustness and generalizability, which is why it is widely used in machine learning training and testing. However, this advantage can actually weaken the result of controlled experiments.

Remark 3.1 (An Objective Perspective on Controlled Experiment Partitioning) *The primary objective of a comparative experiment is to apply different controlled factors to distinct subgroups and assess their effectiveness by comparing the conditional outcome distributions. Therefore, the subsampling process should aim to produce subgroups that closely resemble the original sample's distribution, minimizing the risk of type I and type II errors from imbalanced splits.*

In contrast, the advantage of randomization in statistical tests arises from two key aspects: the simplicity due to the law of large numbers (via subsample size, repeated trials, or both) and the improvement of model robustness and generalizability. Randomized subgroups effectively capture potential differences between the sample data and the true population.

However, such distributional discrepancies weaken the outcomes of controlled experiments. For instance, in hypothesis testing where the null hypothesis posits that the controlled factor has no effect, and the alternative suggests it is effective, any distributional shifts between subgroups must be incorporated into the null distribution. This adjustment narrows the rejection region, reducing the test's power against the alternative hypothesis.

Thus, in controlled experiments where the law of large numbers is less applicable and robustness is not the primary concern, subsampling should prioritize forming subgroups that closely replicate the original sample's distribution. This strategy enhances the statistical power of the experimental results.

3.2 Covariate Adaptive Randomization

A general framework of covariate adaptive randomization aims to design a group assignment strategy that minimizes an imbalance score, which involves the selected covariates and random treatment membership assignment. But there are at least the following disadvantages: (1) manual discretization of continuous covariates, (2) lack of optimality criteria related to comparative test performance guarantee, and (3) a lack of theoretical guarantee of the

similarity among the resulting sensitive groups concerning either the selected covariates or other feature variables in the data set.

For example, both stratified randomization and the basic version of minimization aim to balance the cardinality of the members in subgroups with respective to some covariates. Here, the imbalance score is defined as the difference in the number of members sharing the same (discretized) covariate value across different subgroups.

A more general framework of covariate adaptive randomization replaces the cardinality imbalance score with the following

$$\text{Imbalance} := ||\sum_{i=1}^{n}(2T_i - 1)f(X_i)||^2,$$

where T_i is the random subgroup assignment, X_i denotes the selected covariates, and f is a function to take care of higher order statistics of the covariates. But such a framework still focuses on the difference between the average: such an imbalance score is based on the *difference between the averages* of the subgroups. Therefore, it largely ignores the distributional differences across the subgroups.

In comparison, WHOMP also minimizes an imbalance score. But the score is now an *average of the differences* between the optimally matched or coupled members. By switching the order of difference and averaging by leveraging the coupling (w.r.t. all the feature variables in the data set) technique from optimal transport, the WHOMP objective can capture the distributional differences (w.r.t. all the feature variables in the data set) across the subgroups.

Therefore, WHOMP could also be considered a covariate adaptive randomization technique, albeit one that comes with an optimality criterion that provides provable guarantees for comparative test performance and improved practical outcomes.

3.3 Rerandomization

We will demonstrate that WHOMP can be viewed as a rerandomization method while at the same time addressing two key shortcomings of traditional rerandomization methods. Specifically, traditional rerandomization methods rely on mean discrepancies to quantify covariate imbalances. However, as discussed earlier, mean discrepancies capture only limited aspects of distributional differences, often missing higher-order statistical discrepancies in the data. Additionally, conventional rerandomization methods typically require a manually selected threshold for determining whether a partition is acceptable, which introduces subjectivity.

The following discussion shows that WHOMP resolves the two issues by leveraging optimal transport and unsupervised learning techniques:

- *Comprehensive Imbalance Quantification*: WHOMP utilizes Wasserstein distance to design a covariate imbalance metric that not only accounts for mean differences but also captures a broader range of distributional discrepancies between subgroups. By switching the order of taking differences and averaging, the WHOMP distributional discrepancy metric ensures a more thorough and nuanced assessment of subgroup similarity.

- *Self-learned Threshold*: WHOMP replaces the manual threshold selection with an automated threshold learned by balanced K-means directly from the data based on the requested subgroup number. This self-learned threshold reduces subjectivity and makes the method more objective and consistent across different datasets.

Lemma 3.2 (Equivalence between WHOMP and Rerandomization) *Assume that the balanced K-means problem (Definition 2.2) has a unique solution, denoted by P. Define the subgroup splitting accept and reject rule $\Phi(X, Q)$ by*

$$\Phi(X, Q) := \mathbb{1}_{\{\mathrm{Var}(\bar{X}_Q)=\mathrm{Var}(\mathbb{E}(X_P))\}}(Q). \tag{16}$$

Then WHOMP Random (Algorithm 1) is equivalent to rerandomization with $\Phi(X, Q)$.

Proof See Appendix B ■

3.4 Anti-clustering

We first review anti-clustering, then provide a provable and efficient estimation method for anti-clustering, and finally show the problem of scale difference across subgroups in anti-clustering and how WHOMP solves the problem.

Anti-clustering was first introduced in [37]. The name comes from the fact that the objective is the exact opposite of the classic K-means clustering objective. In particular, given a data set $\{x_i\}_{i=1}^N$, K-means has the following objective

$$\min_{P\in\mathbf{P}(N,K)} \sum_{p\in P}\sum_{x\in X_p} \frac{1}{|p|}||x - \mathbb{E}(X_p)||_2^2 \equiv \min_{P\in\mathbf{P}(N,K)} \sum_{p\in P} \mathrm{Var}(X_p)$$

In contrast, anti-clustering is an optimization problem that has the opposite objective of K-means:

Definition 3.1 (Anti-clustering [37]) *For a fixed $c \in [N]$, the following optimization problem is call anti-clustering with partition cardinality c:*

$$\max_{Q\in\mathbf{P}(N,c)} \sum_{q\in Q}\sum_{x\in X_q} ||x - \mathbb{E}(X_q)||_2^2 \tag{17}$$

The following result shows that, when restricting the partition to have uniform cardinality across subgroups, the expected value of the anti-clustering objective via a random selection across partition elements coincides with the objective of the (balanced) K-means objective:

Lemma 3.3 (Random selection duality for anti-clustering)

$$\min_{\substack{P\in\mathbf{P}(N,K)\\ |p|\equiv c}} \sum_{p\in P} \mathrm{Var}(X_p) \iff \max_{\substack{P\in\mathbf{P}(N,K)\\ |p|\equiv c}} \mathbb{E}_{\mathcal{Q}\sim uniform(\mathcal{Q}(P))}[\sum_{q\in\mathcal{Q}} \mathrm{Var}(X_q)] \tag{18}$$

Proof See Appendix B. ■

In short, if we create a partition $\mathcal{Q} \sim uniform(\mathcal{Q}(P))$ via the random selection method across subgroups in a given P, then the expected anti-clustering objective for the randomly selected $\mathcal{Q}$ (RHS in (35)) is equivalent to the K-means objective for P (LHS in (35)).

For theoretical interest, we also show the following result which is the counterpart of Lemma 3.3 above.

Proposition 3.1 (Random selection duality for clustering)

$$\max_{\substack{Q\in\mathbf{P}(N,c)\\|q|\equiv K}} \sum_{q\in Q} \mathrm{Var}(X_q) \iff \min_{\substack{Q\in\mathbf{P}(N,c)\\|q|\equiv K}} \mathbb{E}_{\mathcal{P}\sim uniform(\mathcal{P}(Q))}[\sum_{p\in\mathcal{P}} \mathrm{Var}(X_p)] \tag{19}$$

Proof See Appendix B ■

Here, the setting is similar to the setting above, except that we fix Q first, then define $\mathcal{P}(Q)$, and generate random partition $\mathcal{P} \sim uniform(\mathcal{P}(Q))$.

Interestingly, as we will show later in Section 4, this random selection from the K-means approach coincides with the optimal solutions to WHOMP. That is, the random estimation of anti-clustering outperforms the exact solution to anti-clustering in terms of the WHOMP objective. Now, we illustrate the disadvantages of the anti-clustering objective and its exact solution compared to WHOMP solutions.

In practice, diversity within each subgroup resulting from partitioning is desirable, as a more diverse subgroup better captures the overall structure of the sample data. This rationale underlies the anti-clustering objective: maximizing the sum of variance within subgroups to promote diversity.

While diversity is advantageous, scale differences across subgroups are undesirable, particularly in mean squared error (MSE) or L^2 loss scenarios. For instance, comparing MSE loss in cross-validation or hold-out sets is problematic when training and test datasets differ by a scale factor, even if their data structures are otherwise similar.

Now, we show that anti-clustering tends to produce subgroups at different scales. To start, we need the following characterization of the anti-clustering objective.

Lemma 3.4 (Centroid Variable Characterization, [37])

$$\max_{P\in\mathbf{P}(N,K)} \sum_{p\in P}\sum_{x\in X_p} ||x - \mathbb{E}(X_p)||_2^2 \iff \min_{P\in\mathbf{P}(N,K)} \sum_{p\in P} |p|||\mathbb{E}(X_p) - \mathbb{E}(X)||_2^2 \tag{20}$$

Intuitively, an anti-clustering partition generates diversity within the resulting subgroups by enforcing low variance among the subgroup centroids. However, the enforcement of low variance among the subgroup centroids leads to scale differences across the resulting subgroups for the following reason: To have similar centroids or means for subgroups, data points sharing the same scale tend to be group together to balance each other so that centroids of subgroups can stay as close as possible.

Now, we use an example to show that anti-clustering leads to diversity in terms of variance or structure across the elements due to the enforced homogeneity of the centroids.

Example 3.1 (Variance scale differences across anti-clustering subgroups) *Given the data points $\{x_i\}_{i\in[9]}$ as shown in Figure 1, the left plot uses the dash lines to connect data points that belong to the same element of an anti-clustering partition. In this case, we have the anti-clustering partition is the following:*

$$\mathcal{P}_{anti-clustering} = \{\{x_1, x_5, x_9\}, \{x_2, x_6, x_7\}, \{x_3, x_4, x_8\}\}.$$

such a partition is guaranteed to be the anti-clustering because the right-hand side of the centroid variable characterization is zero and therefore minimized.

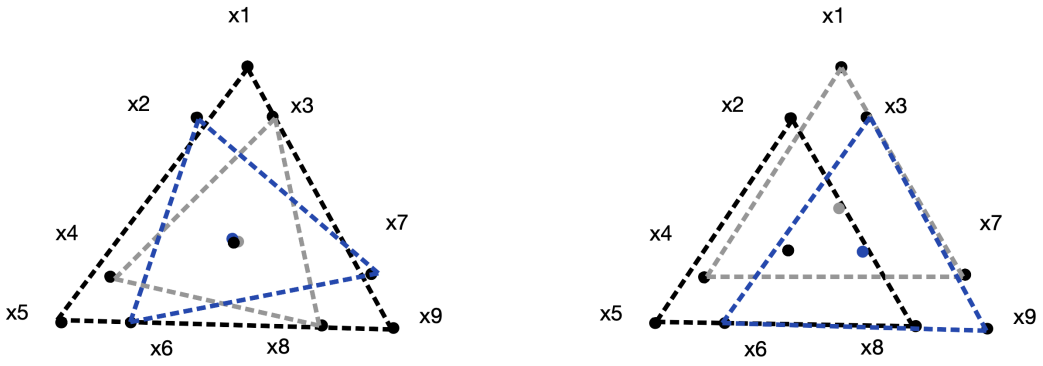

Figure 1: This example illustrates that anti-clustering (left) tends to produce subgroups at different scales: compare the size of the larger triangle formed by x_1, x_5, x_9 with the size of the smaller triangles formed by x_2, x_6, x_7 and x_3, x_4, x_8, respectively. In comparison, WHOMP Matching (right) leads to the desired subgroup partition at the same scale.

On the other hand, if we hope to not only obtain diversity within each of the partition elements but also a similarity in structure, scale, or variance across the elements. Then the desired partition would be the following:

$$\mathcal{P}_{desired} = \{\{x_1, x_4, x_7\}, \{x_2, x_5, x_8\}, \{x_3, x_6, x_9\}\}.$$

As shown in the figure, the right plot uses dash lines to connect data points that belong to the same desirable partition element in this example.

4. Overcoming the NP-hardness of Wasserstein Homogeneity Partition

In this section, we first characterize the set of all optimal solutions to the Wasserstein Homogeneity Partition Problem (WHOMP) by utilizing solutions to the balanced K-means problem (Definition 2.2). Using this characterization, combined with the estimated balanced K-means solution, we propose an efficient approach for computing approximate WHOMP solutions. We also highlight a trade-off between the homogeneity of the mean and variance among the optimal solutions, illustrating how "anti-clustering" and the Wasserstein barycenter represent the two extremes of this trade-off.

Since the balanced K-means is a special case of the 2-norm clustering problem with cardinality constraints, it is known to be NP-hard [5]. Therefore, the provable equivalence between the balanced K-means and WHOMP implies that the proposed WHOMP problem is also NP-hard. To mitigate this complexity, we employ multiple random initialization, the fact that K-means is equivalent to the Wasserstein barycenter problem, and a constrained K-means clustering algorithm inspired by [7] to estimate the balanced K-means solution. It is worth noting that, while our approach uses the constrained K-means algorithm for implementation, any alternative balanced K-means estimation method could be employed instead.

4.1 Characterization of WHOMP Solutions

We first derive a characterization of the solution to (3) and then apply this characterization to design an algorithm to construct the WHOMP partition. To simplify notation in the rest of this section, given a partition $P \in \mathbf{P}(N, c)$, we denote the *Wasserstein barycenter* of $\{X_p\}_{p\in P}$ by $\bar{X}_P := \text{Bary}(\{X_p\}_{p\in P})$. Also, we denote $\mathbb{E}(X_P) = (\mathbb{E}(X_p))_{p\in P}$ to be the vector of the centroids of X_P. Finally, for a partition $P \in \mathbf{P}(N, c)$ we define the partitions resulting from selection from P as follows:

Definition 4.1 (Partition Selected from P: $\mathcal{Q}(P)$) *Given bijective maps $\{T_p\}_p$ where each T_p map $\mathcal{L}(\bar{X}_P)$ to $\mathcal{L}(X_p)$, we construct a partition*

$$Q(\{T_p\}_p) := \{q(\bar{x})\}_{\bar{x}\in\bar{X}_P} \tag{21}$$

with $q(\bar{x}) := \{i : x_i \in \cup_{p\in P}\{T_p(\bar{x})\}$. We define $\mathcal{Q}(P)$ to be the set of all the partitions of the above form:

$$\mathcal{Q}(P) := \bigcup_{\{T_p\}_p : T_p \text{ is bijective}} \{Q(\{T_p\}_p) : T_{p\sharp}\mathcal{L}(\bar{X}_P) = \mathcal{L}(X_p)\} \tag{22}$$

Now, we are ready to state the main result of this section:

Theorem 4.1 (Wasserstein Homogeneity Partition Characterization) *Given P a solution to the K-means partition under the uniform cardinality constraint:*

$$P \in \underset{\substack{P\in\mathbf{P}(N,K),\\ |p|\equiv c}}{\arg\min} \sum_{p\in P} \text{Var}(X_p), \tag{23}$$

then $\mathcal{Q}(P)$ are the set of solutions to (3):

$$\mathcal{Q}(P) = \underset{\substack{Q\in\mathbf{P}(N,c),\\ |q|\equiv K}}{\arg\min} \sum_{q\in Q} \mathcal{W}_2^2(X_q, X). \tag{24}$$

To prove the above result, we need the following two lemmas, which are also of independent interest.

Lemma 4.1 (Maximum Variance Barycenter Characterization of (3))

$$\min_{\substack{Q\in\mathbf{P}(N,c),\\ |q|\equiv K}} \sum_{q\in Q} \mathcal{W}_2^2(X_q, X) \equiv \max_{\substack{Q\in\mathbf{P}(N,c),\\ |q|\equiv K}} \mathrm{Var}(\bar{X}_Q). \tag{25}$$

See proof in Appendix C. That is to say, for any fixed partition cardinality $c \in \mathbb{N}$, the subgroups resulting from a Wasserstein homogeneity partition of cardinality c satisfy that their Wasserstein barycenter has the largest variance among all possible partitions of cardinality c with element-wise uniform cardinality constraint. On the other hand, if one obtains a partition satisfying the Wasserstein barycenter of the resulting subgroups which has a larger variance than any other subgroup barycenters resulting from partitions with cardinality c, then the partition is a solution to the Wasserstein homogeneity partition.

Therefore, to prove Theorem 4.1, it suffices to show that any $Q \in \mathcal{Q}(P)$ results in a $\bar{X}_Q$ with larger or equal variance than any other partition with cardinality c. To that end, we need the following result that relates $\bar{X}_Q$ to $\mathbb{E}(X_P)$:

Lemma 4.2 (Barycenter of X_Q Equals Centroids of X_P for All $Q \in \mathcal{Q}(P)$) *For all $Q \in \mathcal{Q}(P)$, we have*

$$\mathcal{W}_2(\bar{X}_Q, \mathbb{E}(X_P)) = 0. \tag{26}$$

See proof in Appendix C. Finally, we are ready to prove Theorem 4.1 by leveraging the two lemmas above.

Proof [Proof of Theorem 4.1] Assume for contradiction that there is another $Q' \in \mathbf{P}(N, c) \setminus \mathcal{Q}(P)$ such that $q' \equiv K$ and $\mathrm{Var}(\bar{X}_{Q'}) > \mathrm{Var}(\bar{X}_Q)$. Let $T_{q'}$ be the optimal transport maps from $\bar{X}_{Q'}$ to $X_{q'}$ for all $q' \in Q'$. Now, define $p'(\bar{x}) := \{T_{q'}(\bar{x})\}_{q'\in Q'}$ for each $\bar{x} \in \bar{X}_{Q'}$ and $P' := \{p'(\bar{x})\}_{\bar{x}\in\bar{X}_{Q'}}$. It then follows from $|q'| \equiv K$ that $|\bar{X}_{Q'}| = K$. Therefore, we have $P' \in \mathbf{P}(N, K)$ and $|p'(\bar{x})| \equiv \frac{N}{K} = c$ by construction. Also, for each $\bar{x} \in \bar{X}_{Q'}$, we have

$$\mathbb{E}(X_{p'}) = \frac{1}{|Q'|} \sum_{q'\in Q'} T_{q'}(\bar{x}) = \mathrm{Id}(\bar{x}) = \bar{x}. \tag{27}$$

It follows that

$$\mathrm{Var}(\mathbb{E}(X_{P'})) = \mathrm{Var}(\bar{X}_{Q'}) > \mathrm{Var}(\bar{X}_Q) = \mathrm{Var}(\mathbb{E}(X_P)). \tag{28}$$

Here, the last equality follows from Lemma 4.2. But this contradicts the optimality of P. Now, for any $Q, Q' \in \mathcal{Q}(P)$, we have

$$\mathrm{Var}(\bar{X}_{Q'}) = \mathrm{Var}(\mathbb{E}(X_P)) = \mathrm{Var}(\bar{X}_Q). \tag{29}$$

Hence, we have proved by contradiction that each $Q \in \mathcal{Q}(P)$ satisfies $\mathrm{Var}(\bar{X}_Q) \geq \mathrm{Var}(\bar{X}_{Q'})$ for all Q' that satisfy $Q' \in \mathbf{P}(N, c)$ and $q' \equiv K$. Finally, it follows from Lemma 4.1 that Q is a solution to (3). The proof is complete. ■

4.2 Mean and Variance Trade-off among Optimal Solutions

Since the optimal solution to WHOMP is not unique, we are naturally led to the question of how the various optimal solutions differ from each other. To that end, we show a trade-off in variance between the first two moments among the optimal solutions to WHOMP. Furthermore, we show that, among the trade-offs, the extremal solution that minimizes the variance of the subgroup's first moments has the most "anti-clustering" characteristic, while the extremal solution that minimizes the variance of the subgroup's second moments (Algorithm 3) is closely connected to the Wasserstein barycenter; whereas the randomized WHOMP solutions (see details of WHOMP Random design in Algorithm 1) tend to achieve a balance between the two extremes.

Due to the disadvantages of anti-clustering pointed out in Section 3 above, although one can adopt anti-clustering methods to approximate that extremal solution, we focus on the other extreme of the trade-off and balanced solutions in between. Now, we first derive the trade-off.

Lemma 4.3 (Averages variance and variances average trade-off in $\mathcal{Q}(P)$) *Given a partition X_P on the data X with $P \in \mathbf{P}(N, K)$ and $Q \in \mathcal{Q}(P)$, it follows that*

$$\mathrm{Var}(X) = \underbrace{\mathrm{Var}(\mathbb{E}(X_Q))}_{\textit{variance of subgroup expectations}} + \underbrace{\frac{1}{|Q|}\sum_{q\in Q}\mathrm{Var}(X_q)}_{\textit{expected in-subgroup variance}} . \tag{30}$$

Proof It follows directly from law of total variance with $\mathbb{E}(X_Q) = \mathbb{E}(X|Q)$ and

$$\frac{1}{|Q|}\sum_{q\in Q}\mathrm{Var}(X_q) = \mathbb{E}(\mathrm{Var}(X|Q)).$$

■

The above result shows that, when choosing among the optimal solutions to WHOMP, one can choose either the ones resulting in low variance among the subgroup means or the ones with low average subgroup variance.

- *"Anti-clustering"*: On the one hand, it is clear that the solutions that choose to maximize $\frac{1}{|Q|}\sum_{q\in Q}\mathrm{Var}(X_q)$ share the same spirit as anti-clustering. But such maximization of variance often results in scale differences among the subgroups. To see the potential scale problem, Lemma 4.3 shows that maximization of $\frac{1}{|Q|}\sum_{q\in Q}\mathrm{Var}(X_q)$ necessarily leads to minimization of $\mathrm{Var}(\mathbb{E}(X_Q))$. Therefore, in order to enforce low variance in the subgroup averages, the partition tends to group data points sharing the same scale together to achieve balance within the subgroups and have subgroups' averages close to the sample average.

- *Barycenter matching*: On the other hand, if one chooses to minimize the average variance and enforce the scale similarity among the resulting subgroups, Lemma 4.3 shows that it is necessary to increase the variance of the subgroup averages. The

following result shows that the multi-marginal matching in constructing the barycenter of X_P provides a solution that coincides with the extremal solution of the trade-off that minimizes the average variance.

Theorem 4.2 (Barycenter of X_P equals $\mathbb{E}(X_Q)$ with the largest variance in $\mathcal{Q}(P)$) *For all $Q \in \mathcal{Q}(P)$, we have*

$$\mathrm{Var}(\mathbb{E}(X_Q)) \leq \mathrm{Var}(\bar{X}_P). \tag{31}$$

Furthermore, let T_p denote the optimal transport map between $\bar{X}_P$ and X_p for all $p \in P$, then the equality holds as

$$\mathcal{W}_2(\mathbb{E}(X_{Q(\{T_p\}_p)}), \bar{X}_P) = 0. \tag{32}$$

See proof in Appendix C. It is then straightforward to combine the above two results to show that $X_{Q(\{T_p\}_p)}$ is the partition in $\mathcal{Q}(P)$ that minimizes the average variance:

Corollary 4.1 *Let $Q' := Q(\{T_p\}_p)$ as constructed in Theorem 4.2. For all $Q \in \mathcal{Q}(P)$, we have*

$$\frac{1}{|Q|}\sum_{q\in Q'} \mathrm{Var}(X_q) \leq \frac{1}{|Q|}\sum_{q\in Q} \mathrm{Var}(X_q) \tag{33}$$

We finish this section by showing that WHOMP matching, which is the extremal WHOMP solution minimizing expected in-subgroup variance (Algorithm 3), gives the desired solution in Example 3.1.

Remark 4.1 (Homogeneity partition gives desirable subgroups in Example 3.1) *Continuing with Example 3.1, the marginal subgroups can be obtained by performing 3-means on $\{x_i\}_{i\in[9]}$ to obtain*

$$P = \{p_1 = \{x_1, x_2, x_3\}, p_2 = \{x_4, x_5, x_6\}, p_3 = \{x_7, x_8, x_9\}\}.$$

Now, to find the extremal WHOMP solution minimizing expected in-subgroup variance, we first find the Wasserstein barycenter of $\{p_i\}_{i\in[3]}$. Then, for each point on the barycenter, we find the pre-images to form

$$Q = \{q_1 = \{x_1, x_4, x_7\}, q_2 = \{x_2, x_5, x_8\}, q_1 = \{x_3, x_6, x_9\}\},$$

which is exactly the desirable partition. See also Figure 1.

5. Algorithm Design

In this section, we present two algorithms for WHOMP (Definition 2.1) solutions based on our theoretical results in Section 4 and one algorithm that we used to estimate the Balanced K-means (Definition 2.2) solution:

- *WHOMP Random*: Applying randomness to balance the averages' variance (variance of subgroup expectations) and variances' average (expected in-subgroup variance).
- *WHOMP Matching*: Applying Wasserstein matching to minimize the expected in-subgroup variances.
- *Balanced K-means*: Applying optimal transport with Lloyd's algorithm to find K-means solution with uniform cardinality constraint on clusters.

5.1 WHOMP Random:

Algorithm 1: WHOMP Random

Input: sample data set $\{X_i\}_{i=1}^N$;
Step 1: Balanced K-means clustering:;
Obtain balanced K-means clustering (K-means with uniform cardinality constraint) on $\{X_i\}_{i=1}^N$:

$$P := \{p_k\}_{k=1}^K.$$

Step 2: Random selection without replacement;
while $j \in [\frac{N}{K}]$ **do**
  Draw $x'_k \in p_k$ without replacement, for each $k \in [K]$;
  Form $q_i := \{x'_k\}_{k \in K}$;
end
Output: $Q := \{q_i\}_{i=1}^{\frac{N}{K}}$.

WHOMP Random is defined as Algorithm 1. Here, balanced K-means clustering refers to standard K-means clustering with the additional constraint that all resulting clusters must have equal cardinality (i.e., each cluster contains the same number of points). In the implementation used for the numerical experiments in Section 6, we employ Algorithm 2 to estimate the balanced K-means clustering solution. Our algorithm design is inspired by the size-constrained distance clustering [7], and our implementation is inspired by the minimum flow approach to estimate size-constrained K-means.

Algorithm 2: Balanced K-means

Input: sample data set $\{X_i\}_{i=1}^N$, request number of clusters K, max iteration, threshold;
Step 1: Random initialization: Obtain K centers: $\{\bar{x}_i\}_{i=1}^k$, iteration number: iter $= 0$, approximation difference: $\epsilon = \infty$;
Step 2: Iterative Optimal Transport:;
while *iter* $\leq$ *max iteration and* $\epsilon >$ *threshold* **do**
  Find the optimal transport from $\frac{1}{N}\sum_{i=1}^N \delta_{X_i}$ to $\frac{1}{N}\sum_{i=1}^K c\delta_{\bar{x}_i}$, denoted by T;
  Find the pre-images of the optimal transport map for each $\bar{x}_i$: $T^{-1}(\bar{x}_i)$;
  Compute the centroid of the preimages for each $\bar{x}_i$: $\bar{x}'_i := \frac{1}{c}\sum_{x \in T^{-1}(\bar{x}_i)} x$;
  Update $\epsilon = \mathcal{W}_2^2(\frac{1}{K}\sum_{i=1}^K \delta_{\bar{x}_i}, \frac{1}{K}\sum_{i=1}^K \delta_{\bar{x}'_i})$;
  Update iter = iter + 1 ;
  Update $\bar{x}_i = \bar{x}'_i, \forall i \in [K]$
end
Output: $P = \{p_i\}_{i \in [K]}$ where $p_i := T^{-1}(\bar{x}_i), \forall i \in [K]$.

In each of the iteration steps in the Balanced K-means (Algorithm 2), the uniform cardinality constraint is enforced by the uniform weight assigned to each of the centers: If we consider mapping $K*c$ points each with weight $\frac{1}{K*c}$ to K target points each with weight $\frac{1}{K}$ equivalent to mapping $K*c$ points each with weight $\frac{1}{K*c}$ to c copies of the K target points each with weight $\frac{1}{K*c}$, then Choquet's minimization theorem and Birkhoff's theorem together implies that the optimal transport plan is a permutation matrix. Therefore, the optimal transport map assigns c points to each target point.

5.2 WHOMP Matching

We now introduce the WHOMP Matching algorithm (Algorithm 3) to obtain the extremal WHOMP solution that minimizes the expected within-subgroup variance based on Corollary 4.1.

Algorithm 3: WHOMP Matching

Input: sample data set $\{X_i\}_{i=1}^N$: $P := \{p_k\}_{k=1}^K$;
Step 1: Balanced K-means clustering:;
Obtain balanced K-means clustering (K-means with uniform cardinality constraint) on $\{X_i\}_{i=1}^N$:

$$P := \{p_k\}_{k=1}^K.$$

Step 2: Barycenter of K-means clusters:;
Find the Wasserstein barycenter of the clusters in P obtained in Step 1, denoted by

$$\bar{X} := \{\bar{x}_i\}_{i=1}^{\frac{N}{K}},$$

and the corresponding optimal transport map T_k that maps $\bar{X}$ to X_{p_k} for each $k \in [K]$.
Step 3: Group the pre-images of barycenter;
while $i \in [\frac{N}{K}]$ **do**
  Form $q_i := \{T_k(\bar{x}_i)\}_{k\in[K]}$;
end
Output: $Q := \{q_i\}_{i=1}^{\frac{N}{K}}$.

6. Numerical Experiments

The code for the WHOMP (random and matching) implementations, along with the numerical experiments, is available at `https://github.com/xushizhou/WHOMP`.

In this section, we compare the proposed subsampling/partition method, WHOMP, with two baselines: random partitioning and covariate-adaptive randomization (Pocock and Simon's method), using the following datasets:

- Tabular data: synthetic data generated from a Gaussian mixture model

- Tabular data: NPI dataset[2]
- Image data: MNIST [24]
- Graph data: synthetic data generated by a stochastic block model

Before presenting our experimental results, it is important to note that while WHOMP works efficiently for data in moderate or low-dimensional Euclidean spaces, it can also be applied to various data formats by embedding the data into Euclidean space. In this section, we use eigenvectors of the graph Laplacian to embed graph data into Euclidean form. Additionally, we apply t-SNE to embed high-dimensional image data into a lower-dimensional Euclidean space.

6.1 Tabular Data: Gaussian Mixture Model

In this experiment, we test four partition methods: random partitioning, covariate-adaptive randomization (Pocock and Simon's method), WHOMP random, and WHOMP matching, using synthetic data generated from a Gaussian mixture model. To compare the subgroup homogeneity produced by these methods, we perform the following downstream tasks using the subgroups generated by each partitioning method. For all tests, the sample size is fixed at 60 to prevent the law of large numbers. In cases where the law of large numbers applies, one would not expect significant differences between subgroups generated by different partition methods. Additionally, the number of subgroups is fixed at $2, 4, 6$ for all experiments. However, it should be noted that the sample size, number of subgroups, and Gaussian mixture model parameters are all arbitrarily chosen, and we encourage readers to explore WHOMP with different sample sizes and subgroup numbers on other datasets.

6.1.1 Wasserstein-2 Distance Experiment

The goal of this experiment is to validate the theoretical results by comparing the average Wasserstein-2 distance between the subgroups generated by each partition method and the original sample, across 100 repeated tests. A lower average distance and lower variance indicate a better partition method.

Specifically, for each repetition, we begin by randomly drawing 60 data points as the sample dataset, with each 20 points randomly sampled from $\mathcal{N}((0, 10), 3\,\mathrm{Id})$, $\mathcal{N}((-10, -5), 3\,\mathrm{Id})$, and $\mathcal{N}((10, -5), 3\,\mathrm{Id})$. We then apply each partition method to this sample to generate the required number of subgroups. Finally, we compute the average Wasserstein-2 distance between the resulting subgroups and the original sample for each partition method.

Table 1 summarizes the mean and standard deviation of the average Wasserstein-2 distances (between the subgroups and the original sample) across the 100 repetitions.

Figure 2 illustrates the exact distribution of the average Wasserstein-2 distance between the original sample and the resulting subgroups across the 100 repetitions.

2. Raw data from online personality tests: Narcissistic Personality Inventory. Available at the Open-Source Psychometrics Project website: `https://openpsychometrics.org/_rawdata/`

Average (std) W-2 distance between the sample and subgroups			
Partition method	**2 subgroups**	**4 subgroups**	**6 subgroups**
Random	3.481 (0.890)	5.600 (0.849)	6.665 (0.692)
Covariate-adaptive	3.589 (0.939)	5.469 (0.886)	6.566 (0.771)
WHOMP random	1.642 (0.146)	2.575 (0.200)	4.029 (0.211)
WHOMP matching	1.651 (0.145)	2.634 (0.199)	4.170 (0.225)

Table 1: In the table above, we present the mean and standard deviation of the average Wasserstein-2 distances (between the resulting subgroups and the original sample) across the 100 repetitions. The results clearly show that the WHOMP solutions yield both lower average Wasserstein-2 distances and lower standard deviations compared to the random partitioning and covariate-adaptive randomization (Pocock and Simon's method).

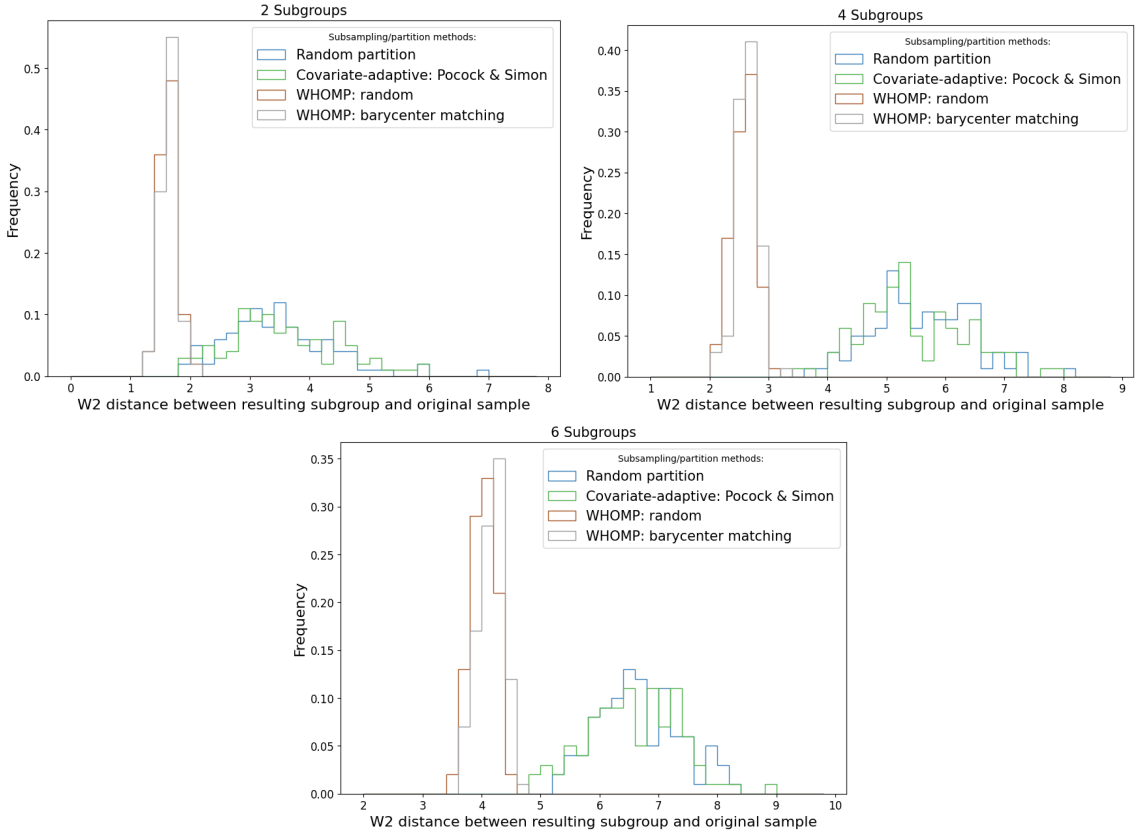

Figure 2: It is evident from the frequency plot above that the worst-case Wasserstein distances resulting from WHOMP solutions are almost the best-case Wasserstein distance resulting from the random partition or Pocock & Simon's covariate-adaptive randomization.

Furthermore, Table 2 presents the variance of the first two moments of the subgroups resulting from the four partition methods. This illustrates the theoretical trade-off, Theorem 4.2, between the variances of the first two moments in the WHOMP solutions.

6.1.2 Classification Experiment: Logistic Regression and SVM

The goal is to assess the distributional homogeneity of the subgroups by using logistic regression or SVM trained on one randomly chosen subgroup to predict the true labels (from the Gaussian mixture model) on another randomly chosen subgroup. For each partition

Variance of subgroups' 1st moments (2nd moments)

Partition method	2 subgroups	4 subgroups	6 subgroups
Random	0.651 (33.376)	2.147 (80.661)	2.990 (134.742)
Covariate-adaptive	0.629 (40.929)	1.787 (80.221)	2.873 (129.885)
WHOMP random	0.093 (23.422)	0.130 (35.381)	0.302 (57.638)
WHOMP matching	0.199 (21.725)	0.555 (27.972)	1.358 (36.988)

Table 2: In the table above, we present the variance of the first and second moments of the subgroups resulting from the different partition methods, averaged over repeated tests where each repetition draws the original sample randomly from Gaussian mixture models. The results reveal a trade-off between the variances of the first and second moments in the WHOMP solutions: WHOMP random exhibits lower variance in the first moment but higher variance in the second moment, while WHOMP matching shows higher variance in the first moment but lower variance in the second moment. Additionally, Pocock and Simon's covariate-adaptive randomization achieves low variance in subgroup averages, aligning with its algorithmic objective. However, it results in a similarly high variance in the second moments as the purely randomized partition method, indicating significant distributional discrepancies.

method, we repeat the prediction test 100 times. Higher average prediction accuracy and lower prediction accuracy variance indicate a better partition method.

For each partition method and repetition, we generate the sample set by randomly selecting 20 data points from $\mathcal{N}((0,10),4\,\mathrm{Id})$, $\mathcal{N}((-10,-5),4\,\mathrm{Id})$, and $\mathcal{N}((10,-5),4\,\mathrm{Id})$. We then apply the partition methods to this sample set to form subgroups, randomly select (without replacement) two of the resulting subgroups as the training and test sets, train a logistic regression model on the training set, and record the test accuracy on the test set.

For the SVM experiment, the goals and test design are identical to those of the logistic regression test, except that logistic regression is replaced with SVM, and the Gaussian mixture model is replaced with $\mathcal{N}((0,10),4\,\mathrm{Id})$, $\mathcal{N}((-10,-5),2\,\mathrm{Id})$, and $\mathcal{N}((10,-5),2\,\mathrm{Id})$.

Table 3 summarizes the outcomes of the logistic regression and SVM experiments.

Logistic Regression Test Accuracy (Standard Deviation)

Partition method	2 subgroups	4 subgroups	6 subgroups
Random	0.981 (0.023)	0.983 (0.032)	0.944 (0.117)
Covariate-adaptive	0.979 (0.017)	0.971 (0.045)	0.950 (0.073)
WHOMP random	0.985 (0.017)	0.982 (0.030)	0.980 (0.040)
WHOMP matching	0.983 (0.017)	0.977 (0.032)	0.984 (0.037)

Support Vector Machine Test Accuracy (Standard Deviation)

Partition method	2 subgroups	4 subgroups	6 subgroups
Random	1.000 (0.000)	0.993 (0.054)	0.987 (0.054)
Covariate-adaptive	1.000 (0.000)	0.998 (0.020)	0.982 (0.061)
WHOMP random	1.000 (0.000)	1.000 (0.000)	0.999 (0.010)
WHOMP matching	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)

Table 3: The table above clearly shows that the WHOMP solutions result in higher average test accuracy with lower standard deviation compared to the other methods. Additionally, the difference becomes more significant as the number of subgroups increases (or equivalently, as the subgroup sizes decrease). Therefore, the advantages of WHOMP are more significant when multiple controlled factors need to be tested.

Linear Regression Test MSE error (Standard Deviation)

Partition method	2 subgroups	4 subgroups	6 subgroups
Random	1.355 (0.112)	1.943 (0.298)	2.548 (0.563)
Covariate-adaptive	1.351 (0.105)	1.988 (0.343)	2.383 (0.370)
WHOMP random	1.291 (0.034)	1.875 (0.142)	2.282 (0.213)
WHOMP matching	1.309 (0.042)	1.861 (0.171)	2.320 (0.230)

Table 4: In the table above, the test mean squared error (MSE) is obtained by training a linear regression model on one randomly chosen subgroup (resulting from the corresponding partition method) and then testing it on another randomly chosen subgroup. The test accuracy average and standard deviation are calculated from 100 repeated tests, with each repetition involving a random draw of the original sample from Gaussian mixture models. It is evident that the WHOMP solutions result in lower MSE and lower standard deviation compared to the random partition and Pocock and Simon's covariate-adaptive randomization methods.

6.1.3 Regression Experiment: Linear Regression

The goal is to evaluate the distributional homogeneity of the subgroups by using a linear regression model trained on one randomly chosen subgroup to predict feature variables on another randomly chosen subgroup. A lower average mean squared error (MSE) and lower error variance indicate a better partition method. More specifically, the test design is the same as the classification experiment, with the distinction that one feature variable in the sample data is chosen as the dependent variable to predict, while the remaining variables serve as independent variables. The results of the experiment are summarized in Table 4.

6.2 Tabular Data: NPI Data Set

We test WHOMP on the NPI dataset, a real-world dataset used in [30] to demonstrate distributional similarity among subgroups. Specifically, for each partition method, we first randomly select 60 data points from the NPI dataset as the sample. We then apply the partition method to generate subgroups and compute the Wasserstein-2 distance between the resulting subgroups and the sample. The sample size is kept small to highlight differences before the Law of Large Numbers affects the results in randomly subsampled data. Each test is repeated 500 times for each partition method. The results of the experiment are summarized in Table 5.

Average (std) W-2 distance between the sample and subgroups

Partition method	2 subgroups	4 subgroups	6 subgroups
Random	2.222 (0.065)	2.808 (0.064)	3.037 (0.063)
Covariate-adaptive	2.227 (0.064)	2.806 (0.064)	3.036 (0.060)
WHOMP random	2.148 (0.055)	2.734 (0.062)	2.965 (0.063)
WHOMP matching	2.179 (0.055)	2.751 (0.062)	2.982 (0.062)

Table 5: In the table above, we present the average and standard deviation of the Wasserstein-2 distance between the resulting subgroups and the original sample, computed over 500 repeated tests for each partition method. Each repetition involves drawing a sample randomly from the NPI dataset. It is evident that the WHOMP solutions achieve both a lower average Wasserstein-2 distance and a lower standard deviation compared to the random partition and Pocock and Simon's covariate-adaptive randomization methods.

Figure 3 shows the exact distribution of the Wasserstein-2 distance between the sample and the resulting subgroups across the 500 tests for the four partition methods.

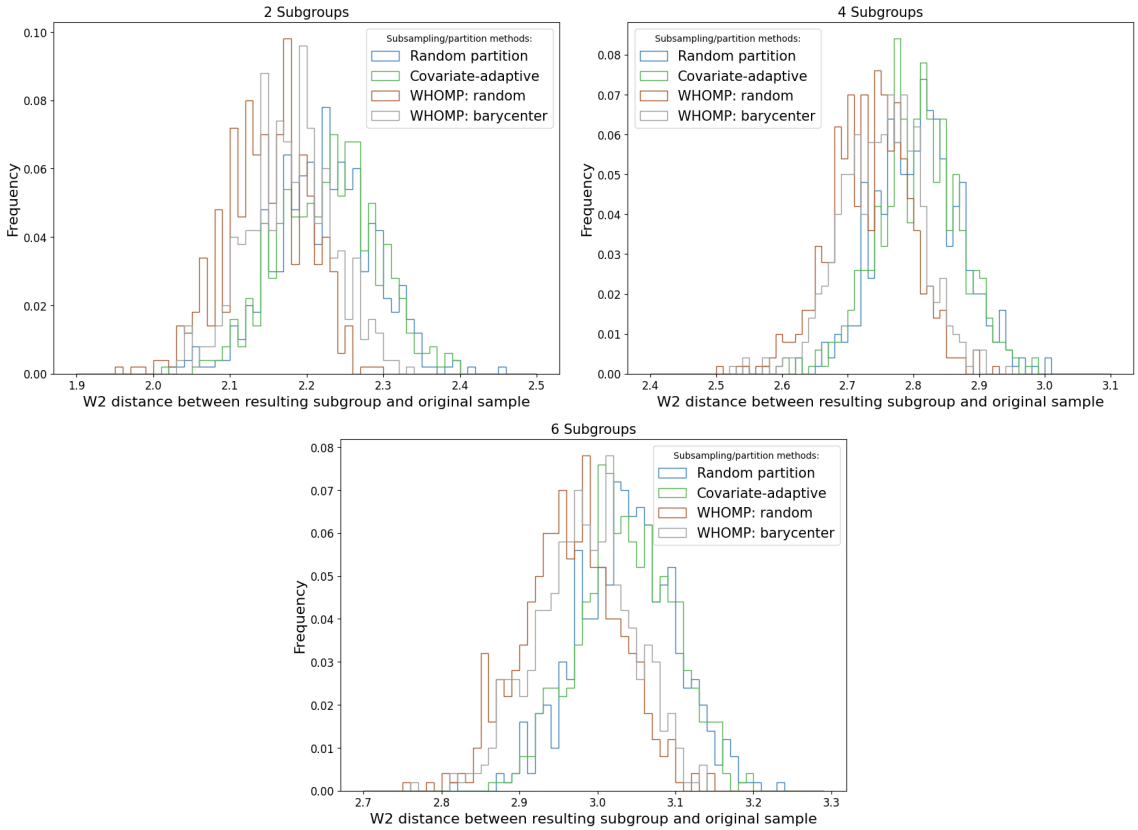

Figure 3: It is evident from the frequency plot above that the worse-case Wasserstein distances resulting from WHOMP solutions are significantly better than the worst-case Wasserstein distance resulting from the random partition or Pocock & Simon's covariate-adaptive randomization. In the case of 2 subgroups, the 90th percentile worst-case distance in WHOMP solutions is equivalent to the 50th percentile worst-case distance in the other two methods.

6.3 Image Data

The goal of this experiment is to demonstrate that WHOMP effectively generates subgroups that are both diverse and homogeneous when applied to image data sets, after composing with embedding methods that embed high-dimensional image data to moderate or low-dimensional Euclidean space. Here, Euclidean closeness should imply closeness in the original image space.

In the experiment, we use t-SNE to embed the MNIST dataset into a 2-dimensional Euclidean space and then apply partition methods to generate partitions on the original image dataset. Due to memory constraints, we reduce the MNIST dataset to 10,000 images in this experiment.

Table 6 presents the mean and standard deviation (over 50 repeated tests) for the normalized entropy of the subgroup frequency vectors. For each subgroup, the frequency vector is defined such that each of the ten entries represents the frequency of images corresponding to a particular digit within the subgroup. Higher normalized entropy values indicate a more balanced (uniformly distributed in terms of digit representation) or more diversified partitioning of the subgroups.

Average (std) of the Subgroup Normalized Entropy in MNIST
Baseline Sample Normalized Entropy = 0.9780

Partition method	2 subgroups	4 subgroups	6 subgroups
Random	0.963 (0.015)	0.945 (0.024)	0.940 (0.026)
Covariate-adaptive	0.963 (0.017)	0.949 (0.017)	0.941 (0.026)
WHOMP random	0.972 (0.007)	0.960 (0.012)	0.959 (0.016)
WHOMP matching	0.973 (0.008)	0.961 (0.014)	0.958 (0.017)

Table 6: The table above shows the average and standard deviation of the normalized entropy of the resulting subgroups, computed over 50 repeated tests for each partition method. It is evident that the WHOMP solutions yield both higher average normalized entropy and lower standard deviation compared to the random partition and Pocock and Simon's covariate-adaptive randomization methods.

6.4 Graph Data

The goal is to demonstrate that WHOMP can be effectively applied to graph data when combined with embedding methods, such as spectral embedding. For each partition method, we perform the following steps and repeat the test 100 times:

1. Generate random graphs from the stochastic block model with three blocks: each has a respective block size of 10, 20, and 30 with a respective edge probability of $[0.6, 0.2, 0.2]$, $[0.2, 0.6, 0.2]$, and $[0.2, 0.2, 0.6]$.
2. Apply spectral embedding to map the graph data into a 2-dimensional Euclidean space.
3. Use the fixed partition method to divide the graph into subgraphs.
4. Compute the Wasserstein-2 distance between the spectrum of the graph Laplacian and the spectrum of the subgraph Laplacian.
5. Compute the average and standard deviation of the Wasserstein-2 distances over all resulting subgraphs.

Table 7 summarizes the mean and standard deviation (over 100 repeated trials) of the single-trial 1st moments (for each trial we compute the average Wasserstein-2 distance between the graph Laplacian spectrum and the resulting subgraph Laplacian spectra.) The goal is to show the expected closeness between the graph and subgraphs in a random single trial and how deviated the closeness is over repeated trials.

Table 8 summarizes the mean and standard deviation (over 100 repeated trials) of the single-trial (square root of) 2nd moments (for each trial we compute the standard deviation of the Wasserstein-2 distance between the graph Laplacian spectrum and the resulting subgraph Laplacian spectra.) The goal is to show the expected stability or uniformity of closeness between the graph and subgraphs in a random single trial and how deviated the uniformity is over repeated trials.

Future Directions

We conclude this paper by briefly sketching some compelling extensions of the WHOMP framework.

Mean (std) of the expected Wasserstein-2 distances between Graph and Subgraph Laplacian Spectra over

Partition method	2 subgroups	4 subgroups	6 subgroups
Random	11.072 (0.307)	16.704 (0.331)	18.525 (0.253)
Covariate-adaptive	11.108 (0.353)	16.689 (0.289)	18.591 (0.293)
WHOMP random	11.275 (0.305)	17.042 (0.262)	18.942 (0.221)
WHOMP matching	11.286 (0.200)	16.563 (0.206)	19.057 (0.249)

Table 7: While achieving nearly the same average expected Wasserstein distance between subgraph and graph Laplacian spectra, WHOMP matching results in lower standard deviation over the 100 trials with randomly sampled graph. That implies WHOMP matching has better stability in partitioning different graphs.

Mean (std) of the Wasserstein-2 distance standard deviation between Graph and Subgraph Laplacian Spectra

Partition method	2 subgroups	4 subgroups	6 subgroups
Random	0.600 (0.429)	0.572 (0.251)	0.583 (0.168)
Covariate-adaptive	0.506 (0.330)	0.645 (0.253)	0.623 (0.219)
WHOMP random	0.398 (0.303)	0.572 (0.241)	0.524 (0.167)
WHOMP matching	0.410 (0.213)	0.455 (0.182)	0.438 (0.136)

Table 8: By achieving lower average (over 100 trials) standard deviation (over subgraphs in each trial) of Wasserstein distance between subgraph and graph Laplacian spectra, WHOMP (especially matching) has better stability in both resulting subgraphs for each trial and partitioning different random graphs over the 50 trials.

- **Sequentially Incoming Data**: Develop algorithms or subgroup assignment mechanisms that optimize the WHOMP objective for sequentially incoming data. This is particularly relevant in scenarios such as clinical trials where participants are enrolled sequentially over time.

- **Cross-Validation**: In essence, the WHOMP objective function quantifies the distributional deviation of the resulting subgroups from the original sample. Thus, by maintaining this function within a specified range (rather than strictly minimizing it, as done in this work), the resulting subgroups can be effectively used for training/testing splits in cross-validation or holdout set generation. This range can be chosen to reflect the realistic distributional variation between the sample and the population distribution.

Acknowledgement

T.S. wants to acknowledge illuminating discussions with Philipp Beineke on the topic of randomized clinical trials. The authors acknowledge support from NSF DMS-2208356, NIH R01HL16351, P41EB032840, and DE-SC0023490.

References

[1] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[2] P. C. Álvarez-Esteban, E. Del Barrio, J. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

[3] A. C. Atkinson. Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika*, 69(1):61–67, 1982.

[4] K. R. Baker and S. G. Powell. Methods for assigning students to groups: A study of alternative objective functions. *Journal of the Operational Research Society*, 53: 397–404, 2002.

[5] A. Bertoni, M. Goldwurm, J. Lin, and F. Saccà. Size constrained distance clustering: separation properties and some complexity results. *Fundamenta Informaticae*, 115(1): 125–139, 2012.

[6] J. Bhadury, E. J. Mighty, and H. Damar. Maximizing workforce diversity in project teams: A network flow approach. *Omega*, 28(2):143–153, 2000.

[7] P. S. Bradley, K. P. Bennett, and A. Demiriz. Constrained k-means clustering. *Microsoft Research, Redmond*, 20(0):0, 2000.

[8] J. Brimberg, N. Mladenović, and D. Urošević. Solving the maximally diverse grouping problem by skewed general variable neighborhood search. *Information Sciences*, 295: 650–675, 2015.

[9] E. Coart, P. Bamps, E. Quinaux, G. Sturbois, E. D. Saad, T. Burzykowski, and M. Buyse. Minimization in randomized clinical trials. *Statistics in Medicine*, 42(28): 5285–5311, 2023.

[10] C. Czado and A. Munk. Assessing the similarity of distributions-finite sample performance of the empirical mallows distance. *Journal of Statistical Computation and Simulation*, 60(4):319–346, 1998.

[11] E. Del Barrio, P. Gordaliza, H. Lescornel, and J.-M. Loubes. Central limit theorem and bootstrap procedure for wasserstein's variations with an application to structural relationships between distributions. *Journal of Multivariate Analysis*, 169:341–362, 2019.

[12] W. Du, D. Xu, X. Wu, and H. Tong. Fairness-aware agnostic federated learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 181–189. SIAM, 2021.

[13] E. Duflo, R. Glennerster, and M. Kremer. Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4:3895–3962, 2007.

[14] B. Efron. Forcing a sequential experiment to be balanced. *Biometrika*, 58(3):403–417, 1971.

[15] Z.-P. Fan, Y. Chen, J. Ma, and S. Zeng. Erratum: A hybrid genetic algorithmic approach to the maximally diverse grouping problem. *Journal of the Operational Research Society*, 62(7):1423–1430, 2011.

[16] T. Feo, O. Goldschmidt, and M. Khellaf. One-half approximation algorithms for the k-partition problem. *Operations Research*, 40(1-supplement-1):S170–S173, 1992.

[17] T. A. Feo and M. Khellaf. A class of bounded approximation algorithms for graph partitioning. *Networks*, 20(2):181–195, 1990.

[18] R. A. Fisher. The arrangement of field experiments. In *Breakthroughs in statistics: Methodology and distribution*, pages 82–91. Springer, 1992.

[19] M. Gallego, M. Laguna, R. Martí, and A. Duarte. Tabu search with strategic oscillation for the maximally diverse grouping problem. *Journal of the Operational Research Society*, 64(5):724–734, 2013.

[20] C. Harshaw, F. Sävje, D. A. Spielman, and P. Zhang. Balancing covariates in randomized experiments with the Gram–Schmidt walk design. *Journal of the American Statistical Association*, pages 1–13, 2024.

[21] Y. Hu and F. Hu. Asymptotic properties of covariate-adaptive randomization. *The Annals of Statistics*, 40(3):1794 – 1815, 2012. doi: 10.1214/12-AOS983. URL `https://doi.org/10.1214/12-AOS983`.

[22] W. N. Kernan, C. M. Viscoli, R. W. Makuch, L. M. Brass, and R. I. Horwitz. Stratified randomization for clinical trials. *Journal of clinical epidemiology*, 52(1):19–26, 1999.

[23] M. S. Krause and K. I. Howard. What random assignment does and does not do. *Journal of Clinical Psychology*, 59(7):751–766, 2003.

[24] Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[25] W. Ma, P. Li, L.-X. Zhang, and F. Hu. A new and unified family of covariate adaptive randomization procedures and their properties. *Journal of the American Statistical Association*, 119(545):151–162, 2024.

[26] M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

[27] K. L. Morgan and D. B. Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263 – 1282, 2012. doi: 10.1214/12-AOS1008. URL `https://doi.org/10.1214/12-AOS1008`.

[28] A. Munk and C. Czado. Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(1):223–241, 1998.

[29] F. A. O'Brien and J. Mingers. A heuristic algorithm for the equitable partitioning problem. *Omega*, 25(2):215–223, 1997.

[30] M. Papenberg. K-plus anticlustering: An improved k-means criterion for maximizing between-group similarity. *British Journal of Mathematical and Statistical Psychology*, 77(1):80–102, 2024.

[31] S. J. Pocock and R. Simon. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, pages 103–115, 1975.

[32] F. J. Rodriguez, M. Lozano, C. García-Martínez, and J. D. González-Barrera. An artificial bee colony algorithm for the maximally diverse grouping problem. *Information Sciences*, 230:183–196, 2013.

[33] W. F. Rosenberger and J. M. Lachin. *Randomization in clinical trials: theory and practice*. John Wiley & Sons, 2015.

[34] W. F. Rosenberger and O. Sverdlov. Handling Covariates in the Design of Clinical Trials. *Statistical Science*, 23(3):404 – 419, 2008. doi: 10.1214/08-STS269. URL `https://doi.org/10.1214/08-STS269`.

[35] N. W. Scott, G. C. McPherson, C. R. Ramsay, and M. K. Campbell. The method of minimization for allocation to clinical trials: a review. *Controlled clinical trials*, 23(6): 662–674, 2002.

[36] S. Senn. Seven myths of randomisation in clinical trials. *Statistics in medicine*, 32(9): 1439–1450, 2013.

[37] H. Späth. Anticlustering: Maximizing the variance criterion. *Control and Cybernetics*, 15(2):213–218, 1986.

[38] D. Sprott and V. Farewell. Randomization in experimental science. *Statistical Papers*, 34:89–94, 1993.

[39] D. R. Taves. Minimization: a new method of assigning patients to treatment and control groups. *Clinical Pharmacology & Therapeutics*, 15(5):443–453, 1974.

[40] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509. URL `https://books.google.com/books?id=hV8o5R7_5tkC`.

[41] C. Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[42] J. Wang, P. Li, and F. Hu. A/B testing in network data with covariate-adaptive randomization. In *International Conference on Machine Learning*, pages 35949–35969. PMLR, 2023.

[43] S. Xu and T. Strohmer. Fair data representation for machine learning at the pareto frontier. *Journal of Machine Learning Research*, 24(331):1–63, 2023. URL `http://jmlr.org/papers/v24/22-0005.html`.

[44] F. Yates. The comparative advantages of systematic and randomized arrangements in the design of agricultural and biological experiments. *Biometrika*, 30(3/4):440–466, 1939.

[45] R. Zaccone, A. Rizzardi, D. Caldarola, M. Ciccone, and B. Caputo. Speeding up heterogeneous federated learning with sequentially trained superclients. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3376–3382. IEEE, 2022.

A. Appendix: Proofs of Results in Section 2

A.1 Proof of Theorem 2.1

Proof Assume $Y(X, c_0) =_d Y(X, c_1)$ for all $c_0, c_1 \in \mathcal{C}$, then we have $\mathcal{W}_2(Y(X, c_0), Y(X, c_1)) = 0$. Now, by the assumption on the WHOMP objective that $\sum_i \mathcal{W}_2^2(X, X_{q_i}) = 0$, we have $\mathcal{W}_2(X, X_{q_i}) = 0, \forall i \in \{0, 1\}$, which further implies

$$\mathcal{W}_2(Y(X_{q_0}, c_0), Y(X, c_1)) = 0 = \mathcal{W}_2(Y(X, c_0), Y(X_{q_1}, c_1)).$$

But it follows from the triangle inequality that

$$\mathcal{W}_2(Y(X_{q_0}, c_0), Y(X_{q_1}, c_1)) = 0,$$

which is equivalent to $Y(X_{q_0}, c_0) =_d Y(X_{q_1}, c_1)$. That completes the proof for the first statement.

Now, we prove the second statement by contraposition. Let $c_0, c_1 \in \mathcal{C}$ be arbitrary and assume $Y(X_{q_0}, c_0) =_d Y(X_{q_1}, c_1)$. Then we have $\mathcal{W}_2(Y(X_{q_0}, c_0), Y(X_{q_1}, c_1)) = 0$. But we also have

$$\mathcal{W}_2(Y(X_{q_0}, c_0), Y(X, c_1)) = 0 = \mathcal{W}_2(Y(X, c_0), Y(X_{q_1}, c_1))$$

from the assumption on the WHOMP objective $\sum_i \mathcal{W}_2^2(X, X_{q_i}) = 0$. It then follows from the triangle inequality that

$$\mathcal{W}_2(Y(X, c_0), Y(X, c_1)) = 0,$$

which further implies that $Y(X, c_0) =_d Y(X, c_1)$. Finally, since our choice of $c_0, c_1 \in \mathcal{C}$ is arbitrary, that completes the proof for the second statement. ■

A.2 Proof of Lemma 2.1

Proof The statement is a direct corollary of the equivalence between WHOMP Random and the rerandomization Lemma 3.2 and [27, Theorem 2.1]. ■

A.3 Proof of Corollary 2.2

Proof Assume $\frac{1}{|Q|}\sum_{q\in Q}\mathcal{W}_2^2(X_q,X)\leq d$, it follows

$$
\begin{aligned}
&(\frac{1}{|Q|}\sum_{q\in Q}\mathcal{W}_2(X_q,X))^2\leq\frac{1}{|Q|^2}\sum_{q\in Q}\mathcal{W}_2^2(X_q,X)\leq d\\
\Longrightarrow&\frac{1}{|Q|}\sum_{q\in Q}\mathcal{W}_2(X_q,X)\leq\sqrt{d}\\
\Longrightarrow&\frac{1}{|Q|}\sum_{q\in Q}\mathcal{W}_1(X_q,X)\leq\sqrt{d}\\
\Longrightarrow&\mathbb{P}(\{\mathcal{W}_1(X_q,X)>\epsilon\})\leq\frac{\sqrt{d}}{\epsilon}\\
\Longrightarrow&\mathbb{P}(\{\sup_{||h||_L\leq 1}|\mathbb{E}(h(X_q))-\mathbb{E}(h(X))|>\epsilon\})\leq\frac{\sqrt{d}}{\epsilon}.\\
\Longrightarrow&\mathbb{P}(\{\sup_{||h||_L\leq L}|\mathbb{E}(\frac{1}{L}h(X_q))-\mathbb{E}(\frac{1}{L}h(X))|>\epsilon\})\leq\frac{\sqrt{d}}{\epsilon}\\
\Longrightarrow&\mathbb{P}(\{\sup_{||h||_L\leq L}|\mathbb{E}(h(X_q))-\mathbb{E}(h(X))|>L\epsilon\})\leq\frac{\sqrt{d}}{\epsilon}\\
\Longrightarrow&\mathbb{P}(\{\sup_{||h||_L\leq L}|\mathbb{E}(h(X_q))-\mathbb{E}(h(X))|>\epsilon\})\leq\frac{L\sqrt{d}}{\epsilon}.
\end{aligned}
$$

Here, the first line follows from Jensen's inequality, the third follows from $\mathcal{W}_1\leq\mathcal{W}_2$, the forth from Markov inequality, and the fifth from the Kantorovich-Rubinstein duality. ■

A.4 Proof of Theorem 2.2

Proof

$$
\begin{aligned}
\mathbb{E}(||\hat{\tau}(Y,\mathcal{Q})-\tau(Y)||_{l^2}^2)&=\mathbb{E}(||\hat{\tau}(Y,\mathcal{Q})||_{l^2}^2)-||\tau(Y)||_{l^2}^2\\
&=\mathbb{E}(||\frac{|\mathcal{Q}|}{n}\sum_{i=1}^{n}(Y_i(q)\mathcal{Q}_i(q)-Y_i(q')\mathcal{Q}_i(q'))||_{l^2}^2)-||\tau(Y)||_{l^2}^2.
\end{aligned}
$$

Here, the first equation follows from the Lemma 2.1, $\tau(Y)=\mathbb{E}(\hat{\tau}(Y,\mathcal{Q}))$. Now, let T_Q be the bijective map from q to q' satisfying:

$$
\mathcal{W}_2^2(X_{\mathcal{Q}(q)},X_{\mathcal{Q}(q')})=\frac{|Q|}{n}\sum_{i\in q}||X_i-X_{T_Q(i)}||_{l^2}^2.
$$

Then it follows that

$$\begin{aligned}
&\mathbb{E}(||\frac{|\mathcal{Q}|}{n}\sum_{i=1}^{n}(Y_i(q)\mathcal{Q}_i(q)-Y_i(q')\mathcal{Q}_i(q'))||_{l^2}^2)\\
=&\mathbb{E}(||\frac{|\mathcal{Q}|}{n}\sum_{i=1}^{n}\mathcal{Q}_i(q)(Y_i(q)-Y_{T_{\mathcal{Q}}(i)}(q'))||_{l^2}^2)\\
\leq&\mathbb{E}(\frac{|\mathcal{Q}|}{n}\sum_{i=1}^{n}||\mathcal{Q}_i(q)(Y_i(q)-Y_{T_{\mathcal{Q}}(i)}(q'))||_{l^2}^2)\\
=&\mathbb{E}(\frac{|\mathcal{Q}|}{n}\sum_{i=1}^{n}Q_i(q)||Y(x_i,q)-Y(x_{T_{\mathcal{Q}}(i)},q')||_{l^2}^2)\\
\leq&\mathbb{E}(\frac{|\mathcal{Q}|}{n}\sum_{i=1}^{n}L^2Q_i(q)||x_i-x_{T_{\mathcal{Q}}(i)}||_{l^2}^2)\\
=&L^2\mathbb{E}(\mathcal{W}_2^2(X_{\mathcal{Q}(q)},X_{\mathcal{Q}(q')})).
\end{aligned}$$

Here, the first equation follows from the definition of $T_{\mathcal{Q}}$, the second from Jensen's inequality, the third from $\mathcal{Q}_i(q)$ being an indicator function, the fourth from the assumption of the uniform Lipschitz property of Y, the fifth from the definition of $T_{\mathcal{Q}}$. Furthermore, we have

$$\begin{aligned}
L^2\mathbb{E}(\mathcal{W}_2^2(X_{\mathcal{Q}(q)},X_{\mathcal{Q}(q')}))\leq&L^2\mathbb{E}\Big(\frac{|\mathcal{Q}|}{n}\sum_{p\in P}\big(\sum_{i=1}^{n}X_iP_i(p)\mathcal{Q}_i(q)-\sum_{i=1}^{n}X_iP_i(p)\mathcal{Q}_i(q')\big)\Big)\\
=&L^2\frac{|\mathcal{Q}|}{n}\sum_{p\in P}\mathbb{E}\Big(\sum_{i=1}^{n}X_iP_i(p)\mathcal{Q}_i(q)-\sum_{i=1}^{n}X_iP_i(p)\mathcal{Q}_i(q')\Big)\\
\text{claim }\rightarrow=&L^2\frac{|\mathcal{Q}|}{n}\sum_{p\in P}(\frac{2|Q|}{|Q|-1}\operatorname{Var}(X_p))\\
=&L^2\frac{2|Q|}{|Q|-1}(\frac{|\mathcal{Q}|}{n}\sum_{p\in P}\operatorname{Var}(X_p))\\
=&L^2\frac{2|Q|}{|Q|-1}[\operatorname{Var}(X)-\operatorname{Var}(\mathbb{E}(X_P))].
\end{aligned}$$

Here, the first line follows from the construction of $\mathcal{Q}(P)$, the third from the claim that we will prove below, and the final from the law of total variance. Now, for any $Q\in\mathcal{Q}(P)$, we have

$$\begin{aligned}
L^2\frac{2|Q|}{|Q|-1}[\operatorname{Var}(X)-\operatorname{Var}(\mathbb{E}(X_P))]=&L^2\frac{2|Q|}{|Q|-1}[\operatorname{Var}(X)-\operatorname{Var}(\mathbb{E}(\bar{X}_Q))]\\
=&L^2\frac{2|Q|}{|Q|-1}\frac{1}{2}\sum_{q\in Q}\mathcal{W}_2^2(X_q,X)\\
=&L^2\frac{|Q|}{|Q|-1}\sum_{q\in Q}\mathcal{W}_2^2(X_q,X),
\end{aligned}$$

where the first line follows from Lemma 4.2, the second from the proof of Lemma 4.1. It remains to prove the claim. Indeed,

$$\begin{aligned}
&\mathbb{E}(\sum_{i=1}^{n} X_i P_i(p)\mathcal{Q}_i(q) - \sum_{i=1}^{n} X_i P_i(p)\mathcal{Q}_i(q')) \\
=&\frac{1}{|Q|}\sum_{x\in X_p}\frac{1}{|Q|-1}\sum_{x'\in X_p\setminus\{x\}}||x-x'||_{l^2}^2 \\
=&\frac{1}{|Q|(|Q|-1)}\sum_{x\in X_p}\sum_{x'\in X_p}||x-x'||_{l^2}^2 \\
=&\frac{1}{|Q|-1}\sum_{x\in X_p}\frac{1}{|Q|}\sum_{x'\in X_p}||x-x'||_{l^2}^2 \\
=&\frac{1}{|Q|-1}\sum_{x\in X_p}(||x-\mathbb{E}(X_p)||_{l^2}^2 + \mathrm{Var}(X_p)) \\
=&\frac{1}{|Q|-1}\sum_{x\in X_p}||x-\mathbb{E}(X_p)||_{l^2}^2 + \frac{1}{|Q|-1}\sum_{x\in X_p}\mathrm{Var}(X_p) \\
=&\frac{|Q|}{|Q|-1}\frac{1}{|Q}\sum_{x\in X_p}||x-\mathbb{E}(X_p)||_{l^2}^2 + \frac{|Q|}{|Q|-1}\mathrm{Var}(X_p) \\
=&\frac{2|Q|}{|Q|-1}\mathrm{Var}(X_p)
\end{aligned}$$

Here, the first equation follows from the construction of $\mathcal{Q}(P)$ and the forth from the fact that $\sum_{x\in X}||y-x||_{l^2}^2 = ||y-\mathbb{E}(X)||_{l^2}^2 + \mathrm{Var}(X)$. This completes the proof. ■

B. Appendix: Proofs of Results in Section 3

B.1 Proof of Lemma 3.1

Proof Assume for contradiction that there exists a $\{x_{s,i}\}_{i\in[K]} =: X_{sample} \in \mathbb{R}^{d\times K}$ such that

$$\mathcal{W}_2^2(X, X_{sample}) < \min_{\substack{P\in\mathbf{P}(N,K)\\ |p|\equiv c}} \frac{1}{K}\sum_{p\in P}\mathrm{Var}(X_p).$$

By Choquet's Minimization Theorem and Birkhoff's Theorem [41], there exist optimal transport maps $\{T_i\}_{i\in[K]}$ such that each T_i maps c points in X to $x_{s,i}$ for each $i \in [K]$. Therefore, the pre-images $T_i^{-1}(x_{s,i})$ satisfy $\bigcup_{i\in[K]} T_i^{-1}(x_{s,i}) = X$, $T_i^{-1}(x_{s,i}) \cap T_j^{-1}(x_{s,j}), \forall i \neq j$ and $|T_i^{-1}(x_{s,i})| = c$ for all $i \in [K]$. Therefore, $X_{P'} := \{X_{p'}\}_{i\in[K]} := \{T_i^{-1}(x_{s,i})\}_{i\in[K]}$ defines a partition on X that satisfies $|p'| = c, \forall p' \in P'$. But, it follows that

$$\frac{1}{K}\sum_{p'\in P'}\mathrm{Var}(X_{p'}) \leq \mathcal{W}_2^2(X, X_{\mathrm{sample}}) < \min_{\substack{P\in\mathbf{P}(N,K)\\ |p|\equiv c}} \frac{1}{K}\sum_{p\in P}\mathrm{Var}(X_p).$$

This contradicts the definition of the right hand side. That completes the proof. ■

B.2 Proof of Lemma 3.2

Proof Since WHOMP Random generates random partitions from $\mathcal{Q}(P)$, it is equivalent to the accept and reject rule $\mathbb{1}_{\mathcal{Q}(P)}(Q)$. Therefore, it suffices to show that $\mathbb{1}_{\mathcal{Q}(P)}(Q) = \Phi(X, Q)$, or equivalently

$$Q \in \mathcal{Q}(P) \iff \mathrm{Var}(\bar{X}_Q) = \mathrm{Var}(\mathbb{E}(X_P)).$$

($\Rightarrow$) First, assume $Q \in \mathcal{Q}(P)$. It then follows from Lemma 4.2 that $\mathrm{Var}(\bar{X}_Q) = \mathrm{Var}(\mathbb{E}(X_P)$. ($\Leftarrow$) Now, assume for contradiction that $Q \notin \mathcal{Q}(P)$ and $\mathrm{Var}(\bar{X}_Q) = \mathrm{Var}(\mathbb{E}(X_P))$. It follows from Theorem 4.2 that $\mathrm{Var}(\mathbb{E}(X_{P'})) = \mathrm{Var}(\bar{X}_Q)$, where $X_{P'} := \{X_{p'}\}_{p' \in P'}$, $X_{p'} := \{T_q(\bar{x}_Q)\}_{q \in Q}$, and T_q is the optimal transport map that maps $\bar{X}_Q$ to X_q for each $q \in Q$. It follows from $Q \notin \mathcal{Q}(P)$ that $P' \neq P$. But that implies $\mathrm{Var}(\mathbb{E}(X_{P'})) = \mathrm{Var}(\mathbb{E}(X_P))$, which contradicts the uniqueness of P. This completes the proof. ■

B.3 Proof of Lemma 3.3

Proof

Pick an arbitrary $P \in \mathbf{P}(N, K)$ that satisfies $|p| \equiv c$. For the left hand side, we have:

$$\begin{aligned}
\sum_{p \in P} \mathrm{Var}(X_p) &= \sum_{p \in P} \Big(\frac{1}{c} \sum_{i \in p} ||x_i - \frac{1}{c} \sum_{j \in p} x_j||_{l^2}^2 \Big) \\
&= \sum_{p \in P} \Big(\frac{1}{c} \sum_{i \in [N]} ||x_i||_{l^2}^2 - \frac{1}{c^2} \sum_{i,j \in p} \langle x_i, x_j \rangle_{l^2} \Big) \\
&= \frac{1}{c} \sum_{i \in [N]} ||x_i||_{l^2}^2 - \frac{1}{c^2} \sum_{p \in P} \sum_{i,j \in p} \langle x_i, x_j \rangle_{l^2} \\
&= \frac{1}{c} \sum_{i \in [N]} ||x_i||_{l^2}^2 - \frac{1}{c^2} \sum_{i,j \in [N]} \langle x_i, x_j \rangle_{l^2} \mathbb{1}_{\{P(i) = P(j)\}}. \qquad (34)
\end{aligned}$$

Now, for the right hand side, we have:

$$\begin{aligned}
\mathbb{E}_{\mathcal{Q}} \Big[\sum_{q \in Q} \mathrm{Var}(X_q) \Big] &= \sum_{Q \in \mathcal{Q}(P)} \mathbb{P}_{\mathcal{Q}}(Q) \Big[\frac{1}{K} \sum_{i \in [N]} ||x_i||_{l^2}^2 - \frac{1}{K^2} \sum_{q \in Q} \sum_{i,j \in q} \langle x_i, x_j \rangle_{l^2} \Big] \\
&= \frac{1}{K} \sum_{i \in [N]} ||x_i||_{l^2}^2 - \frac{1}{K^2} \sum_{Q \in \mathcal{Q}(P)} \mathbb{P}_{\mathcal{Q}}(Q) \Big[\sum_{q \in Q} \sum_{i,j \in q} \langle x_i, x_j \rangle_{l^2} \Big] \\
&= \frac{1}{K} \sum_{i \in [N]} ||x_i||_{l^2}^2 - \frac{1}{K^2} \sum_{Q \in \mathcal{Q}(P)} \mathbb{P}_{\mathcal{Q}}(Q) \Big[\sum_{i,j \in [N]} \langle x_i, x_j \rangle_{l^2} \mathbb{1}_{\{Q(i) = Q(j)\}} \Big] \\
&= \frac{1}{K} \sum_{i \in [N]} ||x_i||_{l^2}^2 - \frac{1}{K^2} \sum_{i,j \in [N]} \langle x_i, x_j \rangle_{l^2} \Big[\sum_{Q \in \mathcal{Q}(P)} \mathbb{P}_{\mathcal{Q}}(Q) \mathbb{1}_{\{Q(i) = Q(j)\}} \Big].
\end{aligned}$$

But

$$\sum_{Q\in\mathcal{Q}(P)} \mathbb{P}_{\mathcal{Q}}(Q)\mathbb{1}_{\{Q(i)=Q(j)\}} = \mathbb{P}_{\mathcal{Q}}(\{\mathcal{Q}(i)=\mathcal{Q}(j)\}|\mathcal{Q}\in\mathcal{Q}(P)) = \begin{cases} 0 & \text{if } P(i)=P(j) \\ \frac{1}{c} & \text{if } P(i)\neq P(j) \end{cases}.$$

Therefore,

$$\begin{aligned} &\mathbb{E}_{\mathcal{Q}}[\sum_{q\in Q}\mathrm{Var}(X_q)] \\ =&\frac{1}{K}\sum_{i\in[N]}||x_i||_{l^2}^2 - \frac{1}{KN}\sum_{i,j\in[N]}\langle x_i,x_j\rangle_{l^2}\mathbb{1}_{\{P(i)\neq P(j)\}} \\ =&\frac{1}{K}\sum_{i\in[N]}||x_i||_{l^2}^2 - \frac{1}{KN}\sum_{i,j\in[N]}\langle x_i,x_j\rangle_{l^2} + \frac{1}{KN}\sum_{i,j\in[N]}\langle x_i,x_j\rangle_{l^2}\mathbb{1}_{\{P(i)=P(j)\}} \end{aligned}$$

Finally, combine the left and right hand sides, we obtain:

$$\min_{\substack{P\in\mathbf{P}(N,K)\\|p|\equiv c}}\sum_{p\in P}\mathrm{Var}(X_p) \iff \max_{\substack{P\in\mathbf{P}(N,K)\\|p|\equiv c}}\mathbb{E}_{\mathcal{Q}\sim uniform(\mathcal{Q}(P))}[\sum_{q\in Q}\mathrm{Var}(X_q)] \tag{35}$$

■

B.4 Proof of Proposition 3.1

Proof For the right hand side, we have

$$\begin{aligned} \mathbb{E}_{\mathcal{P}}[\sum_{p\in\mathcal{P}}\mathrm{Var}(X_p)] &= \sum_{P\in\mathcal{P}(Q)}\mathbb{P}_{\mathcal{P}}(Q)\Big[\frac{1}{c}\sum_{i\in[N]}||x_i||_{l^2}^2 - \frac{1}{c^2}\sum_{p\in P}\sum_{i,j\in q}\langle x_i,x_j\rangle_{l^2}\Big] \\ &= \frac{1}{c}\sum_{i\in[N]}||x_i||_{l^2}^2 - \frac{1}{c^2}\sum_{i,j\in[N]}\langle x_i,x_j\rangle_{l^2}\Big[\sum_{P\in\mathcal{P}(Q)}\mathbb{P}_{\mathcal{P}}(P)\mathbb{1}_{\{P(i)=P(j)\}}\Big] \\ &= \frac{1}{c}\sum_{i\in[N]}||x_i||_{l^2}^2 - \frac{K}{N^2}\sum_{i,j\in[N]}\langle x_i,x_j\rangle_{l^2}\mathbb{1}_{\{Q(i)\neq Q(j)\}}. \end{aligned}$$

Here, the last equality follows from

$$\sum_{P\in\mathcal{P}(Q)}\mathbb{P}_{\mathcal{P}}(P)\mathbb{1}_{\{P(i)=P(j)\}} = \mathbb{P}_{\mathcal{P}}(\{\mathcal{P}(i)=\mathcal{P}(j)\}|\mathcal{P}\in\mathcal{P}(Q)) = \begin{cases} 0 & \text{if } Q(i)=Q(j), \\ \frac{1}{K} & \text{if } Q(i)\neq Q(j). \end{cases}$$

Now, for the left hand side, we have

$$\begin{aligned}\sum_{q\in Q}\mathrm{Var}(X_q) &= \sum_{q\in Q}(\frac{1}{K}\sum_{i\in q}||x_i-\frac{1}{K}\sum_{j\in q}x_j||_{l^2}^2)\\ &= \frac{1}{K}\sum_{i\in[N]}||x_i||_{l^2}^2-\frac{1}{K^2}\sum_{i,j\in[N]}\langle x_i,x_j\rangle_{l^2}\mathbb{1}_{\{Q(i)=Q(j)\}}\\ &= \frac{1}{K}\sum_{i\in[N]}||x_i||_{l^2}^2-\frac{1}{K^2}\sum_{i,j\in[N]}\langle x_i,x_j\rangle_{l^2}+\frac{1}{K^2}\sum_{i,j\in[N]}\langle x_i,x_j\rangle_{l^2}\mathbb{1}_{\{Q(i)\neq Q(j)\}}.\end{aligned}$$

Therefore, it follows from the left and right hand sides that

$$\max_{\substack{Q\in\mathbf{P}(N,c)\\|q|\equiv K}}\sum_{q\in Q}\mathrm{Var}(X_q)\iff\max_{\substack{Q\in\mathbf{P}(N,c)\\|q|\equiv K}}\mathbb{E}_{\mathcal{P}\sim uniform(\mathcal{P}(Q))}[\sum_{p\in\mathcal{P}}\mathrm{Var}(X_p)]\tag{36}$$

■

C. Appendix: Proofs of Results in Section 4

C.1 Proof of Lemma 4.1

Proof The minimum and maximum below are all over the set $\{Q\in\mathbf{P}(N,c):|q|\equiv K\}$.

$$\begin{aligned}\min_{Q\in\mathbf{P}(N,c)}\sum_{q\in Q}\mathcal{W}_2^2(X_q,X) &\equiv \min_{Q\in\mathbf{P}(N,c)}\sum_{q\in Q}\sum_{q'\neq q}\mathcal{W}_2^2(X_q,X_{q'})\\ &\equiv \min_{Q\in\mathbf{P}(N,c)}\sum_{q\in Q}\sum_{q'\in Q}\mathcal{W}_2^2(X_q,X_{q'})\\ &\equiv \min_{Q\in\mathbf{P}(N,c)}\sum_{q\in Q}\mathcal{W}_2^2(X_q,\bar{X}_Q)\\ &\equiv \min_{Q\in\mathbf{P}(N,c)}[\mathrm{Var}(X)-\mathrm{Var}(\bar{X}_Q)]\\ &\equiv \max_{Q\in\mathbf{P}(N,c)}[\mathrm{Var}(\bar{X}_Q)]\end{aligned}$$

Here, the first line follows from the optimality of optimal transport on subsets: $\mathcal{W}_2^2(X_q,X)=\sum_{q'\in Q}\mathcal{W}_2^2(X_q,X_{q'})$, the third line follows from the fact that

$$2\sum_{q\in Q}\mathcal{W}_2^2(X_q,\bar{X}_Q)=\sum_{q\in Q}\sum_{q'\in Q}\mathcal{W}_2^2(X_q,X_{q'}),$$

the forth line follows from the variance reduction formulation: $\sum_{q\in Q}\mathcal{W}_2^2(X_q,\bar{X}_Q)=\mathrm{Var}(X)-\mathrm{Var}(\bar{X}_Q)$, and the last line follows from the fact that $\mathrm{Var}(X)$ is constant. ■

C.2 Proof of Lemma 4.2

Proof Let $Q \in \mathcal{Q}(P)$ be arbitrary. For each $q \in Q$, let T_q denote the optimal transport map that pushes $\mathcal{L}(\mathbb{E}(X_P))$ to $\mathcal{L}(X_q)$. We claim that $T_q(\mathbb{E}(X_p)) = X_{p\cap q}$ for all $p \in P$. It follows that

$$\frac{1}{c}\sum_{q\in Q} T_q(\mathbb{E}(X_p)) = \frac{1}{c}\sum_{q\in Q} X_{p\cap q} = \mathbb{E}(X_p), \forall \mathbb{E}(X_p) \in \mathbb{E}(X_P).$$

It follows from the fixed point characterization of Wasserstein barycenter [2] that

$$\mathcal{W}_2(\mathbb{E}(X_P), \bar{X}_Q) = 0.$$

It remains to prove the claim. Indeed, assume for contradiction that the claim is not true, then there exists $q \in Q$ such that

$$\begin{aligned}\mathcal{W}_2^2(\mathcal{L}(\mathbb{E}(X_P)), \mathcal{L}(X_q)) &= \frac{1}{K}\sum_{p\in P} ||\mathbb{E}(X_p) - T_q(\mathbb{E}(X_p))||^2 \\ &< \frac{1}{K}\sum_{p\in P} ||\mathbb{E}(X_p) - X_{p\cap q}||^2\end{aligned}$$

But then define a new partition $P' \in \mathbf{P}(N, K)$ by $X_{p'} := \bigcup_{q\in Q} T_q(\mathbb{E}(X_p))$ for each $p' \in P'$. It follows that

$$\begin{aligned}\frac{1}{K}\sum_{p'\in P'} \mathrm{Var}(X_{p'}) &= \frac{1}{K}\sum_{p'\in P'} \frac{1}{c}\sum_{q\in Q} ||\mathbb{E}(X_{p'}) - T_q(\mathbb{E}(X_p))||^2 \\ &\leq \frac{1}{K}\sum_{p'\in P'} \frac{1}{c}\sum_{q\in Q} ||\mathbb{E}(X_p) - T_q(\mathbb{E}(X_p))||^2 \\ &< \frac{1}{K}\sum_{p'\in P'} \frac{1}{c}\sum_{q\in Q} ||\mathbb{E}(X_p) - X_{p\cap q}||^2 \\ &= \frac{1}{K}\sum_{p\in P} \mathrm{Var}(X_p).\end{aligned}$$

Here, the first line follows from the definition of P', the second from the fact that Euclidean average is the Fréchet mean, and the third from the assumption. But that contradicts the optimality of P. Therefore, we have proved the claim by contradiction. Finally, since our choice of $Q \in \mathcal{Q}(P)$ is arbitrary, we are done.

■

C.3 Proof of Theorem 4.2

Proof Let T'_p be the bijective map from $\mathbb{E}(X_q)$ to $X_p \cap X_q$. For each $p \in P$, we have

$$\begin{aligned}
\frac{1}{|Q|}\sum_{q\in Q}\mathrm{Var}(X_q) &= \frac{1}{c}\sum_{q\in Q}(\frac{1}{K}\sum_{x\in X_q}||x-\mathbb{E}(X_q)||^2)\\
&= \frac{1}{c}\sum_{q\in Q}(\frac{1}{K}\sum_{p\in P}||T'_p(\mathbb{E}(X_q))-\mathbb{E}(X_q)||^2)\\
&= \frac{1}{K}\sum_{p\in P}(\frac{1}{c}\sum_{q\in Q}||T'_p(\mathbb{E}(X_q))-\mathbb{E}(X_q)||^2)\\
&= \frac{1}{K}\sum_{p\in P}||X_p-\mathbb{E}(X_Q)||_2^2\\
&\geq \frac{1}{K}\sum_{p\in P}\mathcal{W}_2^2(\mathcal{L}(X_p),\mathcal{L}(\bar{X}_P)).
\end{aligned}$$

Here, the second line follows from the definition of T'_p, the penultimate line from the fact that ${T'_p}_\sharp\mathcal{L}(\mathbb{E}(X_Q)) = \mathcal{L}(X_p)$ for all $p \in P$, and the last line follows from the definition of the Wasserstein-2 barycenter. Now, it follows from Lemma 4.3 that

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}(X_Q)) &= \mathrm{Var}(X) - \frac{1}{|Q|}\sum_{q\in Q}\mathrm{Var}(X_q)\\
&\leq \mathrm{Var}(X) - \frac{1}{K}\sum_{p\in P}\mathcal{W}_2^2(\mathcal{L}(X_p),\mathcal{L}(\bar{X}_P))\\
&= \mathrm{Var}(\bar{X}_P),
\end{aligned}$$

where the last line follows from the variance reduction of the Wasserstein-2 barycenter [43, Lemma 5.6].

Finally, when T_p are the optimal transport maps from $\mathcal{L}(\bar{X}_P)$ to $\mathcal{L}(X_p)$, we have

$$\mathbb{E}(X_{q(\bar{x})}) = \frac{1}{K}\sum_{x\in X_{q(\bar{x})}} x = \frac{1}{K}\sum_{p\in P}T_p(\bar{x}) = \bar{x}. \tag{37}$$

Since this is true for all $\bar{x} \in \bar{X}_P$, we have $\mathcal{W}_2^2(\mathbb{E}(X_{Q(\{T_p\}_p)}), \bar{X}_P) = 0$ which implies $\mathrm{Var}(\mathbb{E}(X_{Q(\{T_p\}_p)})) = \mathrm{Var}(\bar{X}_P)$. That completes the proof. ■