Synthetic Principal Component Design: Fast Covariate Balancing with Synthetic Controls

Yiping Lu[1,2] Jiajin Li[3] Lexing Ying[4] Jose Blanchet[3]

[1]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA
[2]Industrial Engineering & Management Science, Northwestern University, Evanston, IL, USA
[3]Department of Management Science and Engineering (MS&E), Stanford University, Stanford, CA, USA
[4]Department of Mathematics, Stanford University, Stanford, CA, USA

Abstract

The optimal design of experiments typically involves solving an NP-hard combinatorial optimization problem. In this paper, we aim to develop a globally convergent and practically efficient optimization algorithm. Specifically, we consider a setting where the pre-treatment outcome data is available and the synthetic control estimator is invoked. The average treatment effect is estimated via the difference between the weighted average outcomes of the treated and control units, where the weights are learned from the observed data. Under this setting, we surprisingly observed that the optimal experimental design problem could be reduced to a so-called *phase synchronization* problem. We solve this problem via a normalized variant of the generalized power method with spectral initialization. On the theoretical side, we establish the first global optimality guarantee for experiment design when pre-treatment data is sampled from certain data-generating processes. Empirically, we conduct extensive experiments to demonstrate the effectiveness of our method on both the US Bureau of Labor Statistics and the Abadie-Diemond-Hainmueller California Smoking Data. In terms of the root mean square error, our algorithm surpasses the random design by a large margin.

1 INTRODUCTION

Estimating the average effects of a binary treatment is one of the main goals of empirical economic and political studies. Randomization in controlled trials is one of the golden rules for estimating average treatment effects (ATE). Suppose the treatment assignment procedure guarantees that the potential outcomes are independent of the treatment status. In that case, a simple difference-in-mean (i.e., average outcomes of the treated and control units) estimator becomes an unbiased estimator of ATE. Nevertheless, a fully randomized experiment may be affected by a significantly high variance in the final estimation. Such variance can be reduced via exploiting the feature information in the observed data. Naturally, we focus on the following question: *can the observed covariates improve the statistical properties of the ATE estimators via experimental designing?* Rubin [2008], Kasy [2016]. This problem is referred to as *covariate balancing*, which restricts the randomization to achieve covariate balance between treatment groups Efron [1971], Morgan and Rubin [2012].

Covariate balancing has been substantially explored in the literature. One approach is applying covariate balancing via an adequately designed propensity score Imai and Ratkovic [2014], Zhao [2019], which requires that we have access to a reasonably large sample of experimental units. However, large samples of experimental units are not always available in practice. Therefore, if only a moderate number of units are accessible to the treatment, Bansal et al. [2018] solve an NP-hard combinatorial optimization problem to balance the empirical covariates. To address this issue, we aim to design a computationally feasible optimal design of experimental studies.

We specifically consider the optimal design of experiments for the synthetic control estimator as in Abadie and Gardeazabal [2003], Abadie et al. [2010, 2015], which becomes an attractive estimation procedure when only a small number of units can be exposed to the experiment. That is, the experiment designer can observe the pre-treatment panel outcome data for a number of units in a number of time periods. Synthetic control compares treated units with a weighted average of untreated units. The weights are determined via empirical fit on the observed pre-treatment outcome. As an example, consider the study of Proposition 99's effect presented in Abadie et al. [2010]. Proposition 99 is a large-scale anti-smoking legislation program that California implemented in 1988. The policy maker wants to estimate the effect of this piece of legislation. Synthetic

Accepted for the 40th Conference on Uncertainty in Artificial Intelligence (UAI 2024).

control suggests the policy maker to estimate the counterfactual outcomes after 1988 by using the observed outcomes of the states without legislative restrictions. Abadie et al. [2010] produces the synthetic control by a combination of Colorado, Connecticut, Montana, Nevada, and Utah and find out that annual per-capita cigarette sales in California were about 26 packs lower than what they would have been by the year 2000. Beyond this application, the method of synthetic controls has been used in many other empirical policy evaluation problems, including legalized prostitution Cunningham and Shah [2018], corporate political connections Acemoglu et al. [2016], taxation Kleven et al. [2013], to just name a few. To realize the benefits offered by synthetic control, Doudchenko et al. [2021], Abadie and Zhao [2021] considers how an optimally designed experiment can help the experiment designer to further reduce the variance. Doudchenko et al. [2021], Abadie and Zhao [2021] propose an optimization approach to select the control group based on the observed pre-treatment outcome. Namely, the choice of the treated units aims to balance the weighted average of treated and untreated covariates. As such, the designer can choose the best non-negative weights. Doudchenko et al. [2021] proves that the underlying optimization problem is still NP-hard and Abadie and Zhao [2021] relaxes the optimization problem into a canonical Quadratic Constraint Quadratic Program (QCQP). Nevertheless, the resulting QCQP is rather computationally demanding and applicable algorithms (e.g., SDR(cf. Semidefinite Programming) Luo et al. [2010]) are not guaranteed to reach a global optimum.

In this paper, we aim to design the first globally convergent optimization algorithm for the weighted covariate balancing formulation Doudchenko et al. [2021]. Moreover, the algorithm is practically efficient. We achieve this by first removing the so-called units cardinality constraint being treated in Doudchenko et al. [2021]. Surprisingly, we find out that this relaxation can be shown to be equivalent to a phase synchronization problem Singer [2011]. Although phase synchronization is still an NP-hard problem Zhang and Huang [2006], many practical algorithms have been recently developed for this nonconvex problem, Bandeira et al. [2017], Boumal [2016], Liu et al. [2017], Zhong and Boumal [2018]. Moreover, as we will argue, this nonconvex problem is polynomial-time solvable under a suitable data generating process (*i.e.* average case complexity). Motivated by this line of research, we propose *Synthetic principal component Design* (SPCD), which optimizes the treatment decision via (a normalized variate of) the generalized power method with spectral initialization Chen et al. [2021]. If we assume that the pre-trement data is sampled from the linear fixed-effect model studied in Abadie et al. [2010], Xu [2017], Athey et al. [2021], Ferman [2021] and invoke the realizable assumptions considered in Abadie and Zhao [2021], we can further establish its *global* optimization guarantee and statistical estimation guarantee for the proposed synthetic control procedure. To the best of our knowledge, this is the first computational paradigm of combinatorial optimization-based experiment design which enjoys a global optimization guarantee.

Notations Let $\mathbb{R}^n$ be the n-dimensional Euclidean space. We write matrices in bold capital letters like $\boldsymbol{A}$, vectors in bold lower-case letters like $\boldsymbol{a}$, and scalars in plain letters. Given a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we use $\sigma_{\max}(\boldsymbol{A})$ or $\|\boldsymbol{A}\|$ to denote its largest singular value (i.e., spectral norm), $\sigma_{\min}(\boldsymbol{A})$ its smallest singular value, and a_{ij} its (i,j)-th element.

Paper organization We organize our paper as follows. In Section 2, we introduce the setup in Doudchenko et al. [2021], Abadie and Zhao [2021] where the authors consider covariate balancing under the synthetic control setting and reformulate it as a phase synchronization problem. In Section 3, we introduce the generalized power method and provide global a optimization guarantee under the linear factor model Abadie and Gardeazabal [2003], Xu [2017], Athey et al. [2021], Ferman [2021] with a realizable assumption Abadie and Zhao [2021]. In Section 4, we apply our method to both simulated and real-world data sets. We end with some closing remarks in Section 5.

1.1 MAIN CONTRIBUTIONS

- We show a surprisingly equivalence between the experiment design with synthetic control Doudchenko et al. [2021], Abadie and Zhao [2021] and phase synchronization problem Singer [2011], Boumal [2016], Liu et al. [2017], where separating the experiment and control group can be transformed to finding the phase of a complex signal. We further reveal the hidden connection between covariate balancing with the smallest eigenvector of the gram matrix and propose a spectral method for fast experiment design.
- We propose a novel normalized version of the generalized power method, which enjoys *global* convergence results under certain generative models. The normalization technique can weak the condition for generative model assumptions to guarantee the global optimiality and also consistently improves the empirical results.
- In terms of the root mean square error, our method empirically surpasses random design by a large margin on both synthetic and real-world datasets. Our performance can even exceed 500000 times of rerandomization over the Abadie-Diamond-Hainmueller smoking legislation data.

2 MATHEMATICAL FORMULATIONS

In this section, we introduce the mathematical formulation of synthetic control (SC) and the optimal experiment design

problem studied in Doudchenko et al. [2021].

Problem setup We aim to estimate the effect of a binary treatment under the panel data setting. Researchers have access to the outcome metric of interest $\boldsymbol{Y} \in \mathbb{R}^{N\times T}$ for N units during T time periods. At time T, researchers are required to execute an experiment by assigning a binary treatment described by $D_i \in \{-1, 1\}, i \in [N]$ based on the observed pre-treatment data. If $D_i = 1$, then a treatment needs to be applied to unit i. After the treatment experiment, furthermore outcomes are observed for additional S time periods $t = T + 1, \cdots, T + S$. During this period, every unit $i \in [N]$ in each time period t is associated with the following two random outcomes: $Y_{it}(-1) = \mu_{it} + e_{it}$, and $Y_{it}(1) = Y_{it}(-1) + \tau_i$, where μ_{it} is the base outcome, $\boldsymbol{\tau}$ is the treatment effect aiming to estimate and e_{it} is the zero mean i.i.d idiosyncratic noise with variance $\text{Var}(\epsilon_{it}) = \sigma$. Once treatment D_i is applied, the experimenter is able to realize $Y_{it} = \frac{(D_i+1)}{2}Y_{it}(1) + \frac{(1-D_i)}{2}Y_{it}(-1)$.

Estimating the treatment effect τ is quite challenging because once we implement a treatment on unit j (*i.e.,* $D_j = 1$) and observe the outcome $Y_{j,T+1}(1)$, then counterfactual outcome $Y_{j,T+1}(-1)$ is not observable. With the pre-treatment observation Y_{iT}, synthetic control literature Abadie et al. [2010], Xu [2017] constructs the counterfactual estimate for a treated unit j from a weighted average of other units' outcomes: $\hat{Y}_{j,T+1}(-1) = \sum_{i:D_i=-1} w_i Y_{i,T+1}$. The weights w_i are learned from the pre-treatment observed data via minimizing $\sum_{t=1}^{T}(Y_{jt} - \sum_{i:D_i=0} w_i Y_{it})^2$. Then the treatment effect of unit j we estimate can be written as $\tau_j = Y_{j,T+1} - \hat{Y}_{j,T+1}(-1)$.

2.1 SYNTHETIC DESIGN

In this subsection, we consider the synthetic design objective function proposed for two-way fixed effect in Doudchenko et al. [2021], Abadie and Zhao [2021] and reveal its hidden connection with the phase synchronization problem. Doudchenko et al. [2021], Abadie and Zhao [2021] aim to design treatment assignments $\{D_i = \pm 1\}_{i=1}^N$ and weights $\{w_i \geq 0\}_{i=1}^N$ for outcome experiments at time $T+1$. For we aims to estimate a two-way fixed effect where the treatment effects are homogeneous, we can consider the *weighted average treatment effect on the treated (wATET)* $\tau = \sum_{i=1}^N \frac{D_i+1}{2} w_i \tau_i$ instead Bottmer et al. [2021]. wATET can be estimated as a difference in weighted means estimator $\hat{\tau} = \sum_{i:D_i=1} w_i Y_{i,T+1} - \sum_{i:D_i=-1} w_i Y_{i,T+1}$ with $\sum_{i:D_i=1} w_i = \sum_{i:D_i=-1} w_i = 1$. Following Doudchenko et al. [2021], upon the outcome model, the mean squared error of the difference-in-weighted-means estimator admits the decomposition

$$\begin{aligned}&\mathbb{E}\left[(\hat{\tau}-\tau)^2 | \{D_i, w_i\}_{i=1}^N\right] \\ &= \underbrace{\left(\sum_{i:D_i=1} w_i\mu_{i,T+1} - \sum_{i:D_i=-1} w_i\mu_{i,T+1}\right)^2}_{\text{weighted covariate balancing}} + \sigma\sum_{i=1}^N w_i^2.\end{aligned} \tag{1}$$

The designer aims to design the experiment with a lowest expected mean square error. Thus, Doudchenko et al. [2021] proposed the following mixed-integer programming for experimenting design with Synthetic Control:

$$\begin{aligned}\min_{\{D_i,w_i\}_{i=1}^n} \ & \frac{1}{T}\sum_{t=1}^T\left(\sum_{i:D_i=1} w_i Y_{it} - \sum_{i:D_i=-1} w_i Y_{it}\right)^2 + \sigma\sum_{i=1}^N w_i^2 \\ \text{s. t.} \quad & w_i \geq 0, D_i \in \{-1, 1\}, \ \forall i \in [N], \\ & \sum_{i:D_i=1} w_i = \sum_{i:D_i=-1} w_i = 1.\end{aligned} \tag{2}$$

Remark 1. *We remove the constraint $\sum_{i:D_i=1} D_i = K$ for a given integer $K \in \mathbb{N}$ in the mixed-integer programming in Doudchenko et al. [2021] mainly as this constraints is empirically proved not critical in Abadie and Zhao [2021]. The NP-hard proof in Doudchenko et al. [2021] depends on the constraint $\sum_{i:D_i=1} D_i = K$. In the following discussion, we will show that the problem is also NP-hard even $\sum_{i:D_i=1} D_i = K$ is removed, as the resulting optimization problem can be reformulated as the ℓ_1-PCA McCoy and Tropp [2011], Wang et al. [2021a] and phase Synchronization Singer [2011], Boumal [2016] problems.*

By making a further simplification of the problem (2), we introduce a change of variable $W_i = w_i D_i$. For $w_i \geq 0$, then $D_i = \text{sgn}(W_i)$ and $w_i = |W_i|$. At the same time, the constraint $\sum_{i:D_i=1} w_i = \sum_{i:D_i=-1} w_i = 1$ is equivalent to $\mathbb{1}^\top W = 0$ and the objective function

$$\frac{1}{T}\sum_{t=1}^T\left(\sum_{i:D_i=1} w_i Y_{it} - \sum_{i:D_i=-1} w_i Y_{it}\right)^2 + \sigma\sum_{i=1}^N w_i^2$$

can be reformulated as $W^\top(\boldsymbol{Y}\boldsymbol{Y}^\top + \lambda I)W$, where $W = [w_1, \cdots, w_N]^\top$ and $\mathbb{1} \in \mathbb{R}^N$ is the all one vector. Thus, (2) can be recast into

$$\min_{W\in\mathbb{R}^n, \mathbb{1}^\top W=0, ||W||_1=1} W^\top(\boldsymbol{Y}\boldsymbol{Y}^\top + \sigma\boldsymbol{I})W. \tag{3}$$

Although the reformulation (3) translates the problem into a compact matrix form, it is still a nonconvex problem due to the constraint $||W||_1 = 1$. To deal with the constraint $\mathbb{1}^\top W = 0$, we add an extra term $\lambda(\mathbb{1}^\top W)^2$ to the objective function, where λ is a pre-defined hyper-parameter. Although this penalty method cannot produce the exact global solution, we can still recover the sign of the global solution (see Theorem 1). Once txhe sign of the global solution is

identified, the remaining effort of computing the magnitude reduces to solving a convex problem

$$\begin{aligned}\arg\min_{\{w_i\}_{i=1}^n} \frac{1}{T}\sum_{t=1}^{T}\left(\sum_{i:y(i)=1} w_i Y_{it} - \sum_{i:y(i)=-1} w_i Y_{it}\right)^2 + \sigma\sum_{i=1}^{N} w_i^2 \\ \text{s.t.} \quad w_i \geq 0,\ \forall i \in [N],\ \sum_{i:y(i)=1} w_i = \sum_{i:y(i)=-1} w_i = 1.\end{aligned} \tag{4}$$

Theorem 1. *For large enough λ, the global solution W^* of (3) satisfies*

$$\operatorname{sgn}(W^*) = \operatorname{sgn}\left(\arg\min_{W\in\mathbb{R}^n, ||W||_1=1} W^\top(\boldsymbol{Y}\boldsymbol{Y}^\top + \sigma\boldsymbol{I} + \lambda\mathbb{1}\mathbb{1}^\top)W\right).$$

The following theorem states that the problem is equivalent to another well-known NP-hard non-convex problem — Phase Synchronization Singer [2011].

Theorem 2 (Equivalence between Synthetic Design and Phase Synchronization)**.** *If $x^* \in \mathbb{R}^n$ is the global solution of $\min_{||x||_1=1} ||Ax||_2^2$ for some matrix $A \in \mathbb{R}^{D\times n}$ $(D > n)$ and the matrix $A^\top A \in \mathbb{R}^{n\times n}$ is invertible, then $y^* = \operatorname{sgn}(x^*)$ is the global solution of $\max_{y\in\{-1,+1\}^n} y^\top((A^\top A)^{-1})^\top y$.*

The proof of Theorem 1 and Theorem 2 is omitted in the main text due to page limit and is shown in Appendix.

Remark 2. *Phase synchronization Singer [2011], Bandeira et al. [2017] aims to recover n phases $z_i = e^{i\theta_i}, i \in [n]$ via solving the following optimization problem*

$$\max_{|x_1|=\cdots=|x_n|=1} \quad x^\top \boldsymbol{C} x, \tag{5}$$

where C_{ij} is the noisy observation of $z_i\bar{z_j} = e^{i(\theta_i-\theta_j)}$. Our problem is symbolically equivalent to (5). However, the data generating process is quite different from the Phase synchronization. The matrix we consider is the inverse of the data matrix. In Appendix, we will show that the design we find is actually the first ℓ_1-principal component McCoy and Tropp [2011], Wang et al. [2021a].

3 ALGORITHM DESCRIPTION

In this section, we propose the generalized power method with spectral initialization to solve our problem. The method is inspired by the lines of its early efforts for phase synchronization problems Journée et al. [2010], Boumal [2016], Liu et al. [2017], Zhong and Boumal [2018].

3.1 GENERALIZED POWER METHODS

Spectral relaxation Singer [2011] is the first simple and efficient approach to solve the phase synchronization problem. Singer [2011] relaxed the N constraints $|x_i| = 1, i \in [N]$ to $||x||_2^2 = n$. Then the solution becomes the leading eigenvector. Liu et al. [2017], Zhong and Boumal [2018] showed that the eigenvector estimator is almost close to the global optima under certain data generating process. Following these works, we take our initial guess of the optimal experiment to be $\operatorname{sgn}(v)$, where v is the smallest eigenvector of the matrix $(\boldsymbol{Y}\boldsymbol{Y}^\top + \alpha I + \lambda\mathbb{1}\mathbb{1}^\top)$ with $\alpha, \lambda > 0$ as two pre-defined hyper-parameters.

To further improve the experiment assignment, we utilize the generalized power method Journée et al. [2010], Luss and Teboulle [2013], which considers the linearization of the objective function at the current point and moves towards a minimizer of this linear function over the non-convex set $\mathcal{C}$. It is also referred as to Frank-Wolfe algorithm for non-convex problems in the literature. Thus, the generalized power method enjoys monotonic improvement ([Boumal, 2016, Lemma 8]). When $g(x)$ is a quadratic function taking the form as $x^\top Ax$, the method will become similar to the power methods for the eigenvector problem. The only difference is the normalization step. The power method normalizes the whole vector but the generalized power method normalizes the individual entries. The generalized power method can be also understood as projected gradient descent Smith [1994]. Indeed, the update

$$y^{t+1} = \operatorname{sgn}[((\frac{1}{\beta}\boldsymbol{Y}\boldsymbol{Y}^\top + \frac{\sigma}{\beta}\boldsymbol{I} + \frac{\alpha}{\beta}\mathbb{1}\mathbb{1}^\top)^{-1} + I)y^t]$$

can be understood as a projection step (sgn) after a gradient descent update with step size $\frac{1}{\beta}$:

$$y^{t+1} = \operatorname{sgn}(y^t + \frac{1}{\beta}((\boldsymbol{Y}\boldsymbol{Y}^\top + \sigma\boldsymbol{I} + \alpha\mathbb{1}\mathbb{1}^\top)^{-1})y^t).$$

Thus the algorithm shares a sufficient ascent condition for each iteration. Our algorithm is called Synthetic principal component Design (SPCD) and it is summarized in Algorithm 1.

Normalized Variant In Boumal [2016], Zhong and Boumal [2018], Liu et al. [2017], the global optimality result is highly dependent on the assumption that the top eigenvector of the iteration matrix lies in $\{-1, 1\}^N$. However, in our setup, the top eigenvector of the iteration matrix is the smallest eigenvector of the covariance matrix which may not be a $\{-1, 1\}^N$ vector. This is also the case appearing in the phase retrieval Candes et al. [2015], Chen and Candes [2015] and the degree corrected stochastic block model Zhao et al. [2012], Jin [2015]. Inspired by the SCORE Jin [2015], Jin et al. [2022] method for degree corrected stochastic block models, we further introduce a normalization step to the generalization power method and call the new algorithm Normalized SPCD (cf. NormSPCD, see (7) for details). We use the diagonal component of the inverse covariance as an estimate of the true normalization component. In Appendix, we show that NormSPCD can be interpreted

Algorithm 1 Synthetic principal component Design

Require: Pre-treatment Observations $Y \in \mathbb{R}^{N\times T}$

Set initial treatment assignment guess through $y^0 = \text{sgn}(v)$, where v is the smallest eigenvector of matrix $(\boldsymbol{Y}\boldsymbol{Y}^\top + \alpha \boldsymbol{I} + \lambda \mathbb{1}\mathbb{1}^\top)$, where α, λ are two pre-defined hyper-parameter.

▷ **Spectral Initialization**

while Converged **do**

Select one of the following two boxes to iterate

For SPCD, update the design via ▷ **Generalized power methods**

$$y^{t+1} = \text{sgn}\left[\left((\boldsymbol{Y}\boldsymbol{Y}^\top + \alpha \boldsymbol{I} + \lambda \mathbb{1}\mathbb{1}^\top)^{-1} + \beta \boldsymbol{I}\right) y^t\right], \tag{6}$$

where β is a pre-defined hyper-parameter.

For NormSPCD, update the design via ▷ **Normalize the inverse covariance matrix**

$$y^{t+1} = \text{sgn}\left[\left[(\boldsymbol{Y}\boldsymbol{Y}^\top + \alpha \boldsymbol{I} + \lambda \mathbb{1}\mathbb{1}^\top)^{-1} + \beta \boldsymbol{I}\right] (y^t/d)\right], \tag{7}$$

where $d = \sqrt{\text{diag}((\boldsymbol{Y}\boldsymbol{Y}^\top + \alpha \boldsymbol{I} + \lambda \mathbb{1}\mathbb{1}^\top)^{-1})}$ and $/$ denotes element-wise divide.

end while

Treat Unit i if $y(i) = -\text{sgn}\left(\sum_{i=1}^N y(i)\right)$ and run the experiment.

▷ **To ensure the size of the treated group is smaller than the control group**

as a Riemannian gradient descent with a specific metric. Empirical results show that it is better than the original GPW in Figure 6b. This normalization technique may be of independent interest in other applications.

3.2 GLOBAL OPTIMALITY GUARANTEE

In this subsection, we provide the global optimization guarantee for the (normalized) generalized power method. Bandeira et al. [2017], Boumal [2016], Liu et al. [2017], Zhong and Boumal [2018] have shown that phase retrieval is globally solvable under certain generative models. We will show that generalized power method can globally converge under certain data generating processes, which are quite different from the ones assumed in the previous works. Following Abadie and Zhao [2021], we consider a realizable linear factor model (also referred to as "interactive fixed-effects model") Abadie et al. [2010], Xu [2017], Athey et al. [2021], Ferman [2021], which has already been commonly employed in the literature as a benchmark model to analyze the properties of synthetic control estimators Amjad et al. [2018], Li [2020]. Recently, Shi et al. [2021] justify the linear assumption from an independent causal mechanism viewpoint. The linear latent factor model is stated in the following assumption.

Assumption 1 (Linear Latent Factor Model Abadie et al. [2010], Xu [2017])**.** *The outcomes are generated via the following linear factor model*

$$\begin{aligned} &Y_{jt} = \delta_t + \frac{D_{jt}+1}{2}\tau + \theta_t^T \mu_j + e_{jt}, \\ &\mathbb{E}[e_{jt}|\delta_t, \mu_j, D_{jt}] = 0, \text{Var}[e_{jt}|\delta_t, \mu_j, D_{jt}] = \sigma. \end{aligned} \tag{8}$$

Here δ_t is the time fixed effect; μ_j is the unobserved common factors; θ_t is a vector of unknown factor loading; e_{jt} is the unobserved i.i.d. idiosyncratic noise; τ is the treatment effect that we aim to estimate and D_{jt} is the $\{-1, 1\}$ variable according to the treatment assignment to unit j at time t. More specifically, in the pre-treatment period, $D_{jt} = -1$ for all $\forall j \in [N], t \in [T]$.

To obtain the global optimality result, we further make the following realizable assumption that there is only one realizable experiment (zero error experiment) in population.

Assumption 2 (Realizable Assumption)**.** *There exists a unique parameter $(w_i, D_i)_{i=1}^n$ that satisfies the following conditions:*

- *D_i are binary treatments,* i.e. $D_i \in \{-1, 1\}$.
- *$w_i \geq 0$ and $\sum_{i=1}^n D_i w_i = 0$.*
- *$||w||_2^2 = N$ and $\epsilon \leq |w_i| \leq \frac{1}{\epsilon}$ for all $\forall i \in [N]$.*
- *The weights will balance the covariates,* i.e. $\sum_{i=1}^n w_i D_i \mu_i = 0$.

Remark 3. *This realizable assumption is similar to Li [2020], [Shi et al., 2021, (5)] and [Abadie and Zhao, 2021,*

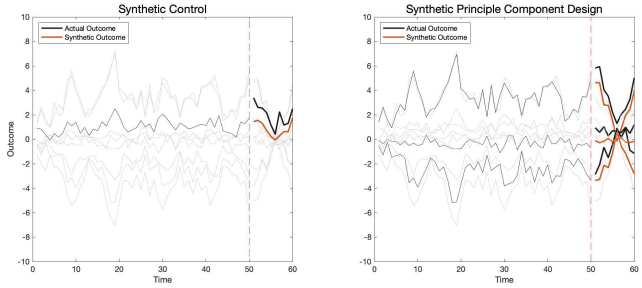


Figure 1: Synthetic principal component design (SPCD) selects the treated units whose features are representative of the whole aggregate market of interest.

Assumption 3]. The difference is that [Abadie and Zhao, 2021, Assumption 3] assumes the weight will cancel noisy observation of the untreated outcome $Y_{jt}(-1)$ which is not realistic when the pre-treatment period T is larger than the number of units N. Our assumption is closer to Li [2020] and [Shi et al., 2021, (5)], but we further assume the uniqueness of the realizable experiment that makes the optimization problem easier (in terms of no need to distinguish different realizable experiments).

Theorem 3. *(Informal) Suppose that Assumptions 1 and 2 hold and that the latent time factor $[\theta_i^\top \delta_i]^\top$ is sampled from a underlying distribution with mean $\tilde{\theta}$ and covariance $\tilde{\Sigma}$. Under regularity assumptions (see Appendix), if σ is small enough and $T \geq \text{poly}(N, \frac{1}{\epsilon})$, then*

- *If $\epsilon > \frac{\sqrt{3}}{2} - 1$, then SPCD converges to the global optima.*
- *If $\epsilon > 0$, then NormSPCD converges to the global optima at a linear rate.*

4 NUMERICAL STUDY

This section reports the numerical tests of our algorithm on both real and simulated data. Figure 1 displays the selected treatment group in the simulated AR(1) process dataset, which demonstrates that our algorithm can identify representative agents from the distribution. Specifically, the results suggest that our algorithm is effective in selecting suitable treatment groups. The result of simulated data is demonstrated in Section 4.1. On the real data, We first follow the same setting as Doudchenko et al. [2021] and then examine our algorithms on two larger datasets. Both experiments have shown the effectiveness of our proposed experimental design algorithm in terms of the root-mean-square error (RMSE), where the squared differences between the true values of the treatment effects and the respective estimates are computed for each treatment period and averaged.

4.1 SIMULATED DATA

We generate data from the linear factor model (also referred to as an "interactive fixed-effects model", Xu [2017], Athey et al. [2021]). The outcome Y comes from

$$Y_{it} = v_t^\top \gamma_i + \tau \frac{D_{it}+1}{2} + e_{it}, \forall i \in [N], \forall t \in [T+S].$$

where γ_i is a vector of latent unit factor of dimension L generated as a standard Gaussian. We simulate the treatment effect τ with ground turth 1. The idiosyncratic noise is sampled from Normal $(0, \sigma^2)$ with $\sigma = 1$. Fixing the test time period $S = 10$ and the number of units $N = 10$, we simulate different pairs of (L, T) selection. We simulated three different choices of the time factor vector v_t

- **Pure Random Latent Vector** In this experiment, we follow Abadie and Zhao [2021] and run our algorithm on a synthetic dataset where all the time latent factors is sampled from random Gaussian. We sample the latent unit factor $v \in \mathbb{R}^{N\times T}$ and latent time factor $\gamma \in \mathbb{R}^{N\times T}$ both as random standard Gaussian matrices. We fix the test time period $S = 10$ and number of units $N = 10$ and simulated different pairs of L, T selection. The final results is shown in Figure 3.
- **Time Varying Factor** In this experiment, we generate the time factor vector v_t as $t - \frac{T+S}{2} + \epsilon_t$, where $t - \frac{T+S}{2}$ is time trend term and e_{it} are i.i.d. standard Gaussian noise. The final results are shown in Figure 2.
- **AR(1) Process** In this experiment, we follow Abadie and Vives-i Bastida [2022] and run our algorithm on a synthetic dataset where the time latent factors is sampled from an AR(1) process. In particular the time factor $\gamma = [\gamma_1, \gamma_2, \cdots, \gamma_T]' \in \mathbb{R}^{N\times T}$ is sampled via
  - $\gamma_1 \sim \mathcal{N}(0, I_N)$,
  - $\gamma_{t+1} = A\gamma_1 + b + \sigma\epsilon, \epsilon \sim \mathcal{N}(0, I_N)$.

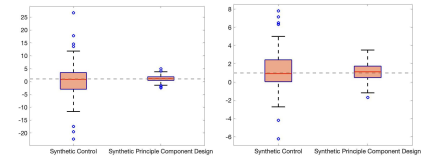
  In our experiment, we take $A = 0.7I_N$, $b = \mathbb{1}$ and $\sigma = 1$. The final result is shown in Figure 4.

Synthetic principale component design (SPCD) selects the treated units whose features are representative of the whole aggregate market of interest Abadie and Zhao [2021]. At the same time, SPCD surpasses SC in every setting in Figure 3, 2 and 4.

4.2 REAL WORLD DATA IN Doudchenko et al. [2021]

To exam our algorithm on real data, we follow Arkhangelsky et al. [2019], Doudchenko et al. [2021] and utilize the US Bureau of Labor Statistics to examine the validity of our algorithm. Besides synthetic control (SC), which randomly selects one unit to implement the treatment, and synthetic design (SD) Doudchenko et al. [2021], we also implement an additional random baseline, which randomly select units as control/group with probability $1/2$. By estimating the average treatment effect on the treated, we compare SC and the random assignment baseline with SPCD in terms of the RMSE. The final result and runing time is shown in Fig 5.

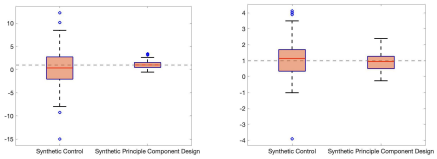
US Bureau of Labor Statistics. We consider a dataset that includes the unemployment rate of 50 states in 40 months



(a) $L = 20, T = 9$ (b) $L = T = 20$

(c) $L = 8, T = 9$ (d) $L = 8, T = 20$

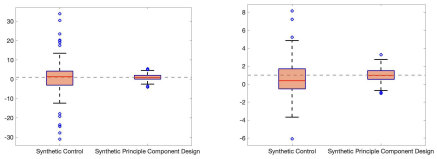
Figure 2: Treatment estimated via synthetic control (SC) and synthetic principale component design (SPCD) over 100 runs of different seeds for different selections of L, T on data generated from time varying time factor.In all cases, Synthetic Principle Design provides more robust estimate of the true treatment effect 1.



(a) $L = 20, T = 9$ (b) $L = 20, T = 20$

(c) $L = 8, T = 9$ (d) $L = 8, T = 20$

Figure 3: Treatment estimated via Synthetic Control and Synthetic Principle Design for data generated from pure random latent vector.



(a) $L = 20, T = 9$ (b) $L = T = 20$

(c) $L = 8, T = 9$ (d) $L = 8, T = 20$

Figure 4: Treatment estimated via Synthetic Control and Synthetic Principle Design for data generated from an AR(1) process.

from the US Bureau of Labor Statistics (BLS). [1] We run 50 simulations such that each simulation utilizes a 20-by-$T+S$ matrix sampled from the original 50-by-40 dataset. More specifically, we randomly select 20 units and use the first T time period to select the synthetic design and synthetic weight. The remaining S time periods are the consecutive months that follow. In our experiment, we fix $S = 5$ and run both experiment for $T = 5, 10$. The final result in terms of the RMSE is shown in Table 1.

4.3 LARGER SCALE REALWORLD DATASET

We further investigated two larger-scale real-world datasets, which have not been simulated in Doudchenko et al. [2021] and Abadie and Vives-i Bastida [2022]. We show our result in Table 1.

The Abadie–Diamond–Hainmueller Smoking Data. Abadie et al. [2010] uses SC to study the effects of Proposition 99, a large-scale anti-smoking legislation program that California implemented in 1988.[2] To simulate the bias of SC and SPCD on this application, following Athey et al. [2021], we consider observations for 38 states (excluding California due to Proposition 99) from 1970 through 2000. We regard the first T year as pre-treatment periods to produce the design and use the last $31 - T$ years as post-treatment periods to test the performance of the treatment assignment. The final result is shown in Table 1 and Figure 6b. Our design surpasses the random design by a large margin on most of the selection of time T. We also compare our method with the rerandomization design Morgan and Rubin [2012], Kallus [2018], Li et al. [2018] in Figure 6c, which shows that our algorithm is still better than 500000 times of rerandomization.

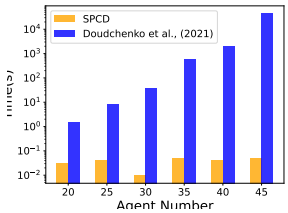
One typical design produced by our algorithm is shown in Figure 6. The plots show that our selection of the control group has the ability to represent all different geographic, demographic, racial, and social structure of states in the United State.

Sales Data The second larger-scale dataset we consider consists of weekly sales of 167 products over 147 weeks,

[1] The BLS Statistics data is available from the BLS website. In this paper, we use the organized version by Arkhangelsky et al. [2019] at https://github.com/synth-inference/synthdid/blob/master/experiments/bdm/data/urate_cps.csv. We thank Arkhangelsky et al. [2019]'s authors carefully organize the data and open source it on github.

[2] The Abadie–Diamond–Hainmueller Smoking data is first organized by Abadie et al. [2010]. In this paper, we use the organized version by Arkhangelsky et al. [2019] at https://github.com/synth-inference/synthdid/blob/master/data/california_prop99.csv which drop the data of minimum wage laws, gun laws to abortion laws in the original data and only considers the smoking outcome data.

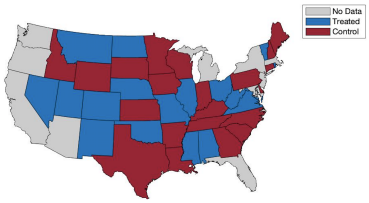
Figure 5: We plot the wall-clock time (seconds) increase with respect to the problem size in the Figure. The wall-clock time for the formulation in Doudchenko et al. [2021] increases exponentially respect to the size of the problem. We also reported the RMSE in the Table. The reported RMSE for BLS dataset are multiplied by 10^3 for readability.



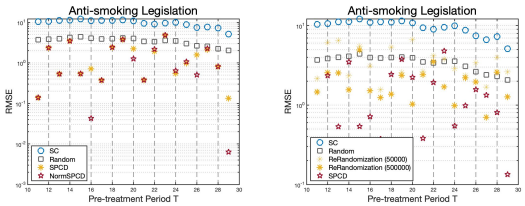
US Bureau of Labor Statistics($N = 20$)							
$T = 5$				$T = 10$			
SC	Random	SD	SPCD	SC	Random	SD	SPCD
14.5	7.5	2.7	**0.9**	11.6	5.6	2.4	**0.6**

Table 1: Root-mean-square errors of the average treatment effect estimates by both synthetic control (SC) and synthetic principal component design (SPCD) on real data. The random design is simulated 10 times and 95% confidence interval is demonstrated.

Sale Data ($N = 147$)						**Anti-smoking legislation** ($N = 40$)					
$T = 100$			$T = 120$			$T = 15$			$T = 25$		
SC	Random	SPCD	SC	Random	SPCD	SC	Random	SPCD	SC	Random	SPCD
2772.1	2183.3	**300.2**	1690.7	658.3	**86.1**	11.65	4.32±0.21	**1.14**	7.89	3.13±0.19	**0.98**



(a) Treatment Selected when $T = 25$



(b) Comparison in terms of RMSE (c) Comparison with rerandomization

Figure 6: A typical design selected via synthetic principal component design (SPCD) and its performance.

collected from a Kaggle competition [3] We regard the first T year as pre-treatment periods to produce the design and use the last $140 - T$ years as post-treatment periods to test the performance of the treatment assignment. The final result is shown in Table 1. Here is an example demonstrating that our algorithm is capable of scaling to larger-sized problems that the previous algorithm Doudchenko et al. [2021] cannot handle.

[3]The Sales data is first organized by https://www.kaggle.com/c/competitive-data-science-predict-future-sales/overview. In this paper, we use the organized version by Farias et al. [2021] at https://github.com/TianyiPeng/Causal-Inference-Code.

5 CONCLUSION

In this paper, we focus on the algorithm design for the optimal experiment design problem for synthetic control proposed in Doudchenko et al. [2021], Abadie and Zhao [2021], which is indeed an NP-hard weighted covariate balancing problem. Surprisingly, the problem can be reduced to a phase synchronization problem and we further propose a fast spectral initialized (normalized) generalized power method to address the resulting optimization problem efficiently. In face of a realizable linear factor model, we provide the first global optimization results for experiment design. Empirically, our method surpasses the original Synthetic Control and random design a large margin in terms of RMSE on various datasets. Additionally, the newly proposed normalization incorporated in GPM may have separate applications in degree correlated stochastic block models.

Acknowledgements

Briefly acknowledge people and organizations here.

All acknowledgements go in this section.

References

Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the basque country. *American economic review*, 93(1):113–132, 2003.

Alberto Abadie and Jaume Vives-i Bastida. Synthetic controls in action. *arXiv preprint arXiv:2203.06279*, 2022.

Alberto Abadie and Jinglong Zhao. Synthetic controls for experimental design. *arXiv preprint arXiv:2108.02196*, 2021.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association*, 105(490):493–505, 2010.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510, 2015.

Daron Acemoglu, Simon Johnson, Amir Kermani, James Kwak, and Todd Mitton. The value of connections in turbulent times: Evidence from the united states. *Journal of Financial Economics*, 121(2):368–391, 2016.

Muhammad Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. *The Journal of Machine Learning Research*, 19(1):802–852, 2018.

Dmitry Arkhangelsky, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. Synthetic difference in differences. Technical report, National Bureau of Economic Research, 2019.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

Afonso S Bandeira, Nicolas Boumal, and Amit Singer. Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Mathematical Programming*, 163(1):145–167, 2017.

Nikhil Bansal, Daniel Dadush, Shashwat Garg, and Shachar Lovett. The gram-schmidt walk: a cure for the banaszczyk blues. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 587–597, 2018.

Lea Bottmer, Guido Imbens, Jann Spiess, and Merrill Warnick. A design-based perspective on synthetic control methods. *arXiv preprint arXiv:2101.09398*, 2021.

Nicolas Boumal. Nonconvex phase synchronization. *SIAM Journal on Optimization*, 26(4):2355–2377, 2016.

Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Yuxin Chen and Emmanuel Candes. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Advances in Neural Information Processing Systems*, 28, 2015.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5): 566–806, 2021.

Scott Cunningham and Manisha Shah. Decriminalizing indoor prostitution: Implications for sexual violence and public health. *The Review of Economic Studies*, 85(3): 1683–1715, 2018.

Nick Doudchenko, Khashayar Khosravi, Jean Pouget-Abadie, Sebastien Lahaie, Miles Lubin, Vahab Mirrokni, Jann Spiess, et al. Synthetic design: An optimization approach to experimental design with synthetic controls. *Advances in Neural Information Processing Systems*, 34, 2021.

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

Vivek Farias, Andrew Li, and Tianyi Peng. Learning treatment effects in panels with general intervention patterns. *Advances in Neural Information Processing Systems*, 34: 14001–14013, 2021.

Bruno Ferman. On the properties of the synthetic control estimator with many periods and many controls. *Journal of the American Statistical Association*, 116(536):1764–1772, 2021.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

Jiashun Jin. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015.

Jiashun Jin, Zheng Tracy Ke, and Shengming Luo. Improvements on score, especially for weak signals. *Sankhya A*, 84(1):127–162, 2022.

Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, 11(2), 2010.

Nathan Kallus. Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):85–112, 2018.

Maximilian Kasy. Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, 24(3):324–338, 2016.

Henrik Jacobsen Kleven, Camille Landais, and Emmanuel Saez. Taxation and international migration of superstars: Evidence from the european football market. *American economic review*, 103(5):1892–1924, 2013.

Nojun Kwak. Principal component analysis based on l1-norm maximization. *IEEE transactions on pattern analysis and machine intelligence*, 30(9):1672–1680, 2008.

Kathleen T Li. Statistical inference for average treatment effects estimated by synthetic control methods. *Journal of the American Statistical Association*, 115(532):2068–2083, 2020.

Xinran Li, Peng Ding, and Donald B Rubin. Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157–9162, 2018.

Huikang Liu, Man-Chung Yue, and Anthony Man-Cho So. On the estimation performance and convergence rate of the generalized power method for phase synchronization. *SIAM Journal on Optimization*, 27(4):2426–2446, 2017.

Zhi-Quan Luo, Wing-Kin Ma, Anthony Man-Cho So, Yinyu Ye, and Shuzhong Zhang. Semidefinite relaxation of quadratic optimization problems. *IEEE Signal Processing Magazine*, 27(3):20–34, 2010.

Ronny Luss and Marc Teboulle. Conditional gradient algorithmsfor rank-one matrix approximations with a sparsity constraint. *siam REVIEW*, 55(1):65–98, 2013.

Michael McCoy and Joel A Tropp. Two proposals for robust pca using semidefinite programming. *Electronic Journal of Statistics*, 5:1123–1160, 2011.

Bamdev Mishra and Rodolphe Sepulchre. Riemannian preconditioning. *SIAM Journal on Optimization*, 26(1):635–660, 2016.

Kari Lock Morgan and Donald B Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282, 2012.

Donald B Rubin. For objective causal inference, design trumps analysis. *The annals of applied statistics*, 2(3): 808–840, 2008.

Claudia Shi, Dhanya Sridhar, Vishal Misra, and David M Blei. On the assumptions of synthetic control methods. *arXiv preprint arXiv:2112.05671*, 2021.

Amit Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, 30(1):20–36, 2011.

Steven T Smith. Optimization techniques on riemannian manifolds. *Fields institute communications*, 3(3):113–135, 1994.

Terence Tao. Topics in random matrix theory. *Graduate Studies in Mathematics*, 132, 2011.

Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.

Peng Wang, Huikang Liu, and Anthony Man-Cho So. Linear convergence of a proximal alternating minimization method with extrapolation for ℓ_1-norm principal component analysis. *arXiv preprint arXiv:2107.07107*, 2021a.

Peng Wang, Zirui Zhou, and Anthony Man-Cho So. Nonconvex exact community recovery in stochastic block model. *Mathematical Programming*, pages 1–37, 2021b.

Stephen Wright, Jorge Nocedal, et al. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76, 2017.

Shuzhong Zhang and Yongwei Huang. Complex quadratic optimization and semidefinite programming. *SIAM Journal on Optimization*, 16(3):871–890, 2006.

Qingyuan Zhao. Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics*, 47(2): 965–993, 2019.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4): 2266–2292, 2012.

Yiqiao Zhong and Nicolas Boumal. Near-optimal bounds for phase synchronization. *SIAM Journal on Optimization*, 28(2):989–1016, 2018.

Synthetic Principal Component Design (Supplementary Material)

Yiping Lu[1,2] Jiajin Li[3] Lexing Ying[4] Jose Blanchet[3]

[1]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA
[2]Industrial Engineering & Management Science, Northwestern University, Evanston, IL, USA
[3]Department of Management Science and Engineering (MS&E), Stanford University, Stanford, CA, USA
[4]Department of Mathematics, Stanford University, Stanford, CA, USA

6 ORGANIZATION OF THE APPENDIX

We organize the appendix as following:

- In Appendix 7, we demonstrated the equivalence between Synthetic Design, ℓ_1-PCA and Phase Synchronization. We also briefly introduce the literature of solving ℓ_1-PCA and Phase Synchronization in Appendix 7.
- In Appendix 8, we demonstrated the proof of global convergence of the generalized power method. The road-map of the proof is following. In Appendix 8.2.1, we analyze the spectral initialization. We showed that it provide as accuracy estimate as the global optima to the ground truth signal. In Appendix 8.2.2, we verify the global optimality of the stationary point of GPW via the Riemann Hessian. In Appendix 8.2.3, we demonstrated the linear convergence rate of the GPW method. In Appendix 8.2.4, we analyze the data generating process to match the assumption needed in the global optimality of the GPW method.

7 EQUIVALENT TO PHASE SYNCHRONIZATION

Theorem 4. *For large enough λ, the global solution W^* of (3) satisfies*

$$\mathrm{sgn}(W^*) = \mathrm{sgn}\left(\underset{W\in\mathbb{R}^n, ||W||_1=1}{\arg\min} W^\top(\boldsymbol{Y}\boldsymbol{Y}^\top + \sigma I + \lambda \mathbb{1}\mathbb{1}^\top)W\right)$$

Proof. We denote $W_\lambda = \arg\min_{W\in\mathbb{R}^n, ||W||_1=1} W^\top(\boldsymbol{Y}\boldsymbol{Y}^\top + \sigma \boldsymbol{I} + \lambda \mathbb{1}\mathbb{1}^\top)W$ for $\lambda > 0$. From [Wright et al., 1999, Theorem 17.1], we know that all the limit point of $W_\lambda \to W^*$ is a global optimum as $\lambda \uparrow \infty$. Thus there exists λ^*, such that for all $\lambda > \lambda^*$, we have

$$||W_\lambda - W^*||_\infty \leq \min\{|W^*(i)| : W^*(i) \neq 0\}.$$

Otherwise there will exist a new limiting point (for $\{x : \|x\| = 1\}$ is compact) which isn't global optimum and contradicted with the previous conclusion. Thus for all $\lambda > \lambda^*$, we have $\mathrm{sgn}(W^*) = \mathrm{sgn}(W_\lambda)$. □

Remark 4. *In the above discussion, we consider both* $\mathrm{sgn}(0) = 1$ *and* $\mathrm{sgn}(0) = -1$ *are right. The reason is that we plug in both sign selection in to the convex programming can both produce the true global optimum.*

Theorem 5 (Equivalence between Synthetic Design, ℓ_1-PCA and Phase Synchronization)**.** *If $x^* \in \mathbb{R}^n$ is the global solution of $\min_{\|x\|_1=1} \|\boldsymbol{A}x\|_2^2$ for some matrix $A \in \mathbb{R}^{D\times n}$ $(D > n)$ and matrix $\boldsymbol{A}^\top\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is invertible, then $y^* = \mathrm{sgn}(x^*)$ is the global solution of $\max_{y\in\{-1,+1\}^n} y^\top((\boldsymbol{A}^\top\boldsymbol{A})^{-1})^\top y$.*

Proof. Firstly, the problem $\min_{\|x\|_1=1} \|\boldsymbol{A}x\|_2^2$ is equivalent to $\min_{\|x\|_1=1} \|(\boldsymbol{A}^\top\boldsymbol{A})^{\frac{1}{2}}x\|_2$ and can be further reduced to $\max_{\|y\|_2=1} \|(\boldsymbol{A}^\top\boldsymbol{A})^{-\frac{1}{2}}y\|_1$ whose optimal solution enjoys the one to one mapping with the original one, that is $y^* = \frac{(\boldsymbol{A}^\top\boldsymbol{A})^{\frac{1}{2}}x^*}{\|(\boldsymbol{A}^\top\boldsymbol{A})^{\frac{1}{2}}x^*\|_2}$. Here, $y^* = \arg\max_{\|y\|_2=1} \|(\boldsymbol{A}^\top\boldsymbol{A})^{-\frac{1}{2}}y\|_1$ and $x^* = \arg\min_{\|x\|_1=1} \|(\boldsymbol{A}^\top\boldsymbol{A})^{\frac{1}{2}}x\|_2$.

Accepted for the 40th Conference on Uncertainty in Artificial Intelligence (UAI 2024).

At the same time, for any matrix $T \in \mathbb{R}^{n\times n}$, we have

$$\max_{\|x\|_2=1} \|\boldsymbol{T}x\|_1 = \max_{\|x\|_2=1, y\in\{-1,+1\}} y^\top \boldsymbol{T}x = \max_{y\in\{-1,+1\}} \|\boldsymbol{T}^\top y\|_2 = \max_{y\in\{-1,+1\}} y^\top \boldsymbol{T}\boldsymbol{T}^\top y. \tag{9}$$

and thus leads to $\arg\max_{y\in\{-1,+1\}} ||\boldsymbol{T}^\top y||_2 = \text{sgn}(\boldsymbol{T}x^*)$ where $x^* = \arg\max_{||x||_2=1} ||\boldsymbol{T}x||_1$. Combining the two facts, we can prove the theorem. □

Remark 5. *Although we formulated the mixed integer programming as a well-known compact matrix form, the two problems,* i.e. *ℓ_1-PCA and phase synchronization, are still known to be NP-hard McCoy and Tropp [2011], Zhang and Huang [2006]. However phase synchronization can be globally solved under certain data generative models Bandeira et al. [2017], Boumal [2016], Liu et al. [2017], Zhong and Boumal [2018]. As far as the author known, there is still no data generative models for ℓ_1-PCA been found can be globally solved. Wang et al. [2021a] show that for the Kurdyka-Lojasiewicz exponent of the ℓ_1-PCA problem at any of the limiting critical points is $\frac{1}{2}$. This allows one to establish the linear convergence to the local stationary point of certain algorithm. Although, Generalized Power Method is also proposed for ℓ_1-PCA Kwak [2008], but only local convergence is guaranteed.*

8 OPTIMIZATION THEORY

Out theory mostly follows Bandeira et al. [2017], Boumal [2016]. But we have slightly different optimization problem (optimization over $\mathbb{C}^n$ in Bandeira et al. [2017], Boumal [2016] but $\mathbb{R}^n$ in ours) and uses different data generating process (gram matrix in Bandeira et al. [2017], Boumal [2016] and inverse of gram matrix in our paper. All the entries of ground truth vector norm equals to 1 in Bandeira et al. [2017], Boumal [2016], *i.e.* $|z_i| = 1$. But this is not assumed in our paper.). For completeness, we complete all the proof details here in the appendix.

8.1 PRELIMINARIES

In this section, we present some basics of Riemannian gradients. For $\{-1,1\}^n$ is a degenerate manifold. In the proof, we will consider the global optimality of the synchronization problem over a larger space $\mathbb{T}^n = \{z \in \mathcal{C}^n : |z_1| = \cdots = |z_n| = 1\}$. Next we endow $\mathbb{T}^n$ with Euclidean metric $\langle y^1, y^2\rangle = \sum_{i=1}^n \mathcal{R}\{y_i^1 y_i^{2H}\}$ which is the equivalent to viewing $\mathcal{C}^n$ as $\mathbb{R}^{2n}$ and equip with the canonical inner product. Then $\mathbb{T}^n$ can be considered as a sub-manifold and the tangent space can be written as

$$\mathcal{T}_y\mathbb{T}^n = \{\dot{y} \in \mathcal{C}^n : \mathcal{R}(\dot{y}_i y_i^H) = 0, \forall i \in [n]\}.$$

The projector to the tangent space is $\text{Proj}_x : \mathbb{C}^n \to T_x\mathbb{T}^n : \dot{x} \to \dot{x} - \mathcal{R}\{\text{ddiag}(\dot{x}x^H)\}x$, where $\text{ddiag} : \mathbb{C}^n \to \mathbb{C}^n$ is a function set all off-diagonal entries of the input matrix to zero. Thus the Riemannian gradient of function $f(x) = x^H\boldsymbol{C}x$ is given as

$$\text{grad} f(x) = 2(\mathcal{R}\{\text{ddiag}(\boldsymbol{C}xx^H)\} - \boldsymbol{C})x$$

Following Bandeira et al. [2017], we consider the Riemannian Hessian on the tangent space as the second-order necessary optimality condition. The Riemannian Hessian is defined as

$$\text{Hess} f(x)[\dot{x}] = \text{Proj}_x \text{Dgrad} g(x)[\dot{x}] = \text{Proj}_x 2S(x)\dot{x},$$

where $S(x) = \mathcal{R}\{\text{ddiag}(\boldsymbol{C}xx^H)\} - \boldsymbol{C}$. If x is a (local) optimum, then $\langle \dot{x}, S(x)\dot{x}\rangle > 0$ for all $\dot{x} \in \mathcal{T}_x\mathbb{T}^N$.

For NormSPCD iteration 7, we consider the update as a Riemannian steepest-descentMishra and Sepulchre [2016]. The Riemannian steepest-descent *search* direction to minimize objective function f as $\arg\min_{\xi_x\in\mathbb{R}^n} \langle \nabla f(x), \xi_x\rangle_{\mathcal{R}} + \frac{1}{2}\langle \xi_x, \xi_x\rangle_{\mathcal{R}}$. The Riemannian metric we consider for NormSPCD on $\mathcal{T}_x\mathbb{T}^N$ defined as

$$\langle y_1, y_2\rangle_{\mathcal{R}} = \sum_{i=1}^N |z_i|^2 \mathcal{R}\left(y_1(i) y_2(i)^H\right) \quad , \forall y_1, y_2 \in \mathcal{T}_x\mathbb{T}^N.$$

Similarly, we can define the new Riemannian Hessian as $S_{\mathcal{R}}(x) = \mathcal{R}\{\text{ddiag}(\mathring{\boldsymbol{C}}xx^H)\} - \mathring{\boldsymbol{C}}$, where $\mathring{\boldsymbol{C}} = \text{diag}(\frac{1}{|z|})C\text{diag}(\frac{1}{|z|})$. We'll show that $rS(x) \preccurlyeq S_{\mathcal{R}}(x) \preccurlyeq RS(x)$ for some constant $r, R > 0$

In our discussion, we consider our algorithm works in the Field of complexity numbers. However, from the closeness of the Field of real numbers, we know that the whole trajectory of our algorithm lies in the Field of real numbers. Global minimum in the complex domain is a harder problem and directly indicate the global optimality in $\{-1,+1\}^N$.

8.2 GLOBAL OPTIMALITY OF (NORMALZIED) GENERALIZED POWER METHODS

In this section, we first study a meta version of the optimization problem. Then we will show how our generative model can be fitted into this framework. We consider the following meta optimization problem

$$\min_{x\in\mathbb{T}^n} f(x) = x^H \boldsymbol{C} x \tag{10}$$

where $\boldsymbol{C} = zz^H + \boldsymbol{\Delta}$ is a Hermite perturbed rank-1 matrix. Different from Bandeira et al. [2017], Boumal [2016], Liu et al. [2017], Zhong and Boumal [2018] which assumes $z \in \mathbb{T}^N$, instead, we have the following assumption on the ground truth vector $z \in \mathbb{R}^N$:

Assumption 3. *For some $\epsilon > 1$, we have*

$$\epsilon \le |z_i| \le \frac{1}{\epsilon}, \qquad \forall i \in [N]$$

This is a smooth optimization problem over a smooth Riemannian manifold $\mathcal{T}^n$. Then the Riemannian gradient $\text{grad} g(x) = 2(\mathcal{R}(\text{diag}(Cxx^H)) - C)x$. The first order necessary optimality condition is $\text{gard} g(x) = 0$. We will also make use the second order optimality via the Riemannian Hessian

$$\text{Hess} g(x)[\dot{x}] = 2\left\langle \dot{x}, \boldsymbol{S}\dot{x} \right\rangle, \forall \dot{x} \in \mathcal{T}_y\mathbb{T}^n$$

where $\boldsymbol{S}(x) = \mathcal{R}(\text{ddiag}(Cxx^H) - C)$ and $\text{ddiag} : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ zeros out all off-diagonal entries of a matrix. Although computing the global optimum of (10) is NP-hard Zhang and Huang [2006], fortunately, global optimality of (10) can sometimes be certified through the Hermitian Hessian matrix $S(x) = \mathcal{R}(\text{ddiag}(Cxx^H) - C)$. This can be shown in the following lemma for sufficient optimality condition:

Lemma 1 (Optimality Gap)**.** *Let x^* be globally optimal for (10). For any $x \in \mathbb{T}^N$, the optimality gap at x can be bounded as*

$$0 \le f(x^*) - f(x) \le -N\lambda_{\min}(\boldsymbol{S}(x)).$$

As a result, if $\boldsymbol{S}(x) \succeq 0$, then x is the global optimality problem for (10).

Proof. See [Bandeira et al., 2017, Section 4.2] and [Boumal, 2016, Lemma 2].

□

In the following lemma, we showed that similar property holds for the changed Riemannian metric.

Lemma 2 (Optimality Gap for Riemannian Formulation)**.** *Let x^* be globally optimal for (10). If $\epsilon < |z_i| < \frac{1}{\epsilon}$, then for any $x \in \mathbb{T}^N$, the optimality gap at x can be bounded as*

$$0 \le f(x^*) - f(x) \le -\frac{1}{\epsilon}N\lambda_{\min}(\boldsymbol{S}_{\mathcal{R}}(x)).$$

As a result, if $\boldsymbol{S}(x) \succeq 0$, then x is the global optimality problem for (10).

Proof. This is because

$$x^H\boldsymbol{C}x - y^H\boldsymbol{C}y = y^H\boldsymbol{S}(x)y \ge \frac{N}{\epsilon}\lambda_{\min}(\boldsymbol{S}_{\mathcal{R}}(x))$$

□

To solve this problem, we consider the following Generalized Power Method Boumal [2016], Liu et al. [2017] in Algorithm 2 and our normalized version in Algorithm 3.

Algorithm 2 Generalized Power Method

Set initialization through $x^0 = \text{sgn}(v)$, where v is the leading eigenvector of matrix C. ▷ Spectral Initialization
Define $\tilde{\boldsymbol{C}} = \boldsymbol{C} + \alpha I_N$ where $\alpha = ||\boldsymbol{\Delta}||$
while Converged **do**
  $x^{t+1} = T(x^t) \triangleq \text{sgn}\left[\tilde{\boldsymbol{C}} x^t\right]$ ▷ Generalized power methods
end while

Algorithm 3 Normalized Generalized Power Method

Set initialization through $x^0 = \text{sgn}(v)$, where v is the leading eigenvector of matrix C. ▷ Spectral Initialization
Define $\tilde{\boldsymbol{C}} = \boldsymbol{C} + \alpha I_N$ where $\alpha = ||\boldsymbol{\Delta}||$
while Converged **do**
  $x^{t+1} = \mathring{T}(x^t) \triangleq \text{sgn}\left[\tilde{\boldsymbol{C}}(x^t ./ \sqrt{\text{ddiag}(\boldsymbol{C})})\right]$, ▷ Normalized Generalized power methods
  where $./$ is the element-wise division.
end while

8.2.1 The Spectral Initialization

We make the initial guess in Algorithm 2 via spectral relaxation Singer [2011]. Denote v as the leading eigenvector of matrix C. From the following Lemma 3, we can know that the leading eigenvector v is close to the ground truth signal z.

Lemma 3. *Given vector $z \in \mathbb{R}^N$ satisfies $||z||_2 = 1$. For matrix $\boldsymbol{C} = zz^\top + \boldsymbol{\Delta}$, where $\boldsymbol{\Delta} \in \mathbb{R}^{N\times N}$ is a symmetric perturbation matrix. Then for all $x \in \mathcal{C}^n$ and $||x||_2^2 = N$ satisfies $x^H \boldsymbol{C} x \geq z^H \boldsymbol{C} z$, we have*

$$\left|\left| \min_{\theta\in\{1,-1\}} \theta x - z \right|\right| \leq \frac{4||\boldsymbol{\Delta}||}{\sqrt{N}}$$

where the $||\boldsymbol{\Delta}||$ is matrix operator norm.

Proof. See [Bandeira et al., 2017, Lemma 4.1] and[Boumal, 2016, Lemma 1]. □

Based on the top eigenvector v, we project v to the Riemann manifold $\mathbb{T}^n$ and make the initial guess. For $\boldsymbol{C}$ is a symmetric real matrix, the eigenvector $v \in \mathbb{R}^N$. Thus projection to $\mathbb{T}^n$ of vector v will simply become $\text{sgn}(v)$. In the next lemma, we'll show the Spectral Estimator is almost as close to z as the global optima.

Lemma 4 (The Spectral Estimator is almost as accuracy as the Global Optima)**.**

$$\left|\left| \min_{\theta\in\{1,-1\}} \theta\text{sgn}(v) - \theta\text{sgn}(z) \right|\right| \leq \frac{8||\boldsymbol{\Delta}||}{\epsilon\sqrt{N}}$$

Lemma 4 is the direct corollary of Lemma 3 and the following technical lemma, which is also important in the convergence rate proof in Section 8.2.3.

Lemma 5. *For $w \in \mathbb{R}^n$ and $z \in \mathbb{R}^n$ satisfies $||z||_2^2 = N$ and $\epsilon \leq |z_i| \leq \frac{1}{\epsilon}, \forall i \in [N]$ (or $1+\epsilon$), then we have*

$$||\text{sgn}(w) - \text{sgn}(z)||_2 \leq \frac{2}{\epsilon}||w - z||_2.$$

Proof. [Zhong and Boumal, 2018, Lemma 13] and [Wang et al., 2021b, Lemma 3] □

8.2.2 The Generalized Power Method

Although, the Spectral Estimator produce good estimates. We still cannot obtain the global optimum of (10). Following Boumal [2016], Liu et al. [2017], we proceed the Generalized Power Method (GPM) to further improve the estimate. Boumal [2016] showed that the Generalized Power Method will converge to the global optima of problem (10) and Liu et al. [2017]

showed that the proceeded estimate is always better than the initial spectral estimate. The procedure of the Generalized Power Method is shown in Algorithm 2. We also consider the Normalized GPM (Algorithm 3) in this section.

For the simplicity of description, we define an equivalence relationship $\sim$ over $\mathbb{T}^n$ as

$$x \sim y \quad \Longleftrightarrow \quad x = ye^{i\theta} \quad \text{for some } \theta \in \mathbb{R}.$$

The quotient space $\mathbb{T}^n/\sim$ is defined as all the corresponding equivalence class $\{xe^{i\theta} : \theta \in \mathbb{R}\}$ for some $x \in \mathbb{C}$. The error measure we are interested in

$$d_q(z, x) = \min_{\theta\in\mathbb{R}} ||xe^{i\theta} - z||_q = \sqrt{2(n - |z^H x|)}, \quad q \in [1, \infty].$$

Lemma 6. *For all $x, y \in \{-1, 1\}^N$ and $q \in [1, \infty]$, then we have*

$$e^{i \arg\min_{\theta\in\mathbb{R}} ||xe^{i\theta}-z||_q} \in \{-1, 1\}.$$

Proof. We use proof by contradiction to prove this statement. If $\theta^* = \arg\min_{\theta\in\mathbb{R}} e^{i||xe^{i\theta}-z||_q} \notin \mathbb{R}$, then we will have

$$||x\text{sgn}(\mathcal{R}(e^{i\theta^*})) - z||_q \le ||x(\mathcal{R}(e^{i\theta^*})) - z||_q < e^{i||xe^{i\theta}-z||_q}.$$

This is contradicted with $\theta^* = \arg\min_{\theta\in\mathbb{R}} e^{i||xe^{i\theta}-z||_q}$. Thus $e^{i \arg\min_{\theta\in\mathbb{R}} ||xe^{i\theta}-z||_q} \in \{-1, 1\}$.

□

Notice that (normalized) GPM iterates on the quotient space $\mathcal{T}^N/\sim$, *i.e.* if $x \sim y$, then $\text{sgn}(\tilde{\boldsymbol{C}}x) \sim \text{sgn}(\tilde{\boldsymbol{C}}y)$. Thus without further notice, all the equality in the following discussion is equality in the quotient, *i.e.* $x = y$ means $x \sim y$.

Lemma 7 (Monotonic Cost Improvement for GPM)**.** *The iterates $\{x^k\}_{k\in\mathbb{N}}$ produced by Algorithm 2 satisfies $f(x^{k+1}) > f(x^k)$ unless converged. Thus the iterates x^k do not cycle.*

Proof. See [Boumal, 2016, Lemma 8] □

Although Algorithm 3 does not guarantee the Monotonic Cost Improvement on the original target function f. We can still prove that the produced iterates from Algorithm 3 do not cycle for it's monotonically improve another energy function.

Lemma 8 (Converging of Normalized GPM)**.** *The iterates $\{x^k\}_{k\in\mathbb{N}}$ produced by Algorithm 3 do not cycle.*

Proof. Consider the potential function $\mathring{f}(x) = \frac{1}{2}x^H \left(\text{diag}(\frac{1}{|z|})\boldsymbol{C}\text{diag}(\frac{1}{|z|})\right) x$. Then the normalized power iteration can be considered as the frank-wolf algorithm for the potential function in the sense that

$$x^{t+1} = \mathring{T}(x^t) = \arg\max_{y\in\mathbb{T}^N} \left\langle y, \left(\text{diag}(\frac{1}{|z|})\boldsymbol{C}\text{diag}(\frac{1}{|z|})\right) x^t \right\rangle.$$

Similar with [Boumal, 2016, Lemma 8], we knows that the iterates $\{x^k\}_{k\in\mathbb{N}}$ monotonically improve the potential function $\mathring{f}$ and thus the iterates do not cycle. □

Lemma 9. *If x is a fixed point of Generalized Power Methods (Algorithm 2), at least one of the following holds*

$$|z^H x| \ge \epsilon N - 4(||\boldsymbol{\Delta}|| + \alpha) \qquad \text{or} \qquad |z^H x| \le \frac{4(||\boldsymbol{\Delta}|| + \alpha)}{\epsilon}.$$

Furthermore, if $||\boldsymbol{\Delta}|| \le \frac{\epsilon^2 N}{13}$ and $\alpha < ||\boldsymbol{\Delta}||$, all the accumulation points x of Algorithm 2 satisfies $|z^H x| \ge \epsilon n - 8||\boldsymbol{\Delta}||$.

Proof. The fixed point of the generalized power method satisfies $(\tilde{\boldsymbol{C}}x)_i\bar{x}_i = |(\tilde{\boldsymbol{C}}x)_i|$. Thus we have

$$x^H\tilde{\boldsymbol{C}}x = ||\tilde{\boldsymbol{C}}x||_1$$

On one hand, the quadratic term $x^H\tilde{\boldsymbol{C}}x$ can be upper bounded as

$$x^H\tilde{\boldsymbol{C}}x = |z^Hx|^2 + x^H\boldsymbol{\Delta}x + \alpha n \le |z^Hx|^2 + (||\boldsymbol{\Delta}|| + \alpha)n.$$

On the other hand $||\tilde{\boldsymbol{C}}x||_1$ can be lower bounded via

$$||\tilde{\boldsymbol{C}}x||_1 = \sum_{i=1}^{N} \left|(z^Hx)z_i + (\boldsymbol{\Delta}x)_i + \alpha x_i\right| \ge N\epsilon|z^Hx| - ||\boldsymbol{\Delta}x||_1 - \alpha N.$$

At the same time, $||\boldsymbol{\Delta}x||_1 \le \sqrt{N}||\boldsymbol{\Delta}x||_2 \le N||\boldsymbol{\Delta}||$. Combine this with the two previous inequality, we get

$$|z^Hx|(\epsilon N - |z^Hx|) \le 2N(||\boldsymbol{\Delta}|| + \alpha).$$

The above inequality enforces that one of $|z^Hx| \ge \epsilon N - 4(||\boldsymbol{\Delta}|| + \alpha)$ and $|z^Hx| \le \frac{4(||\boldsymbol{\Delta}|| + \alpha)}{\epsilon}$. holds. We call all the stationary point satisfies $|z^Hx| \ge \epsilon N - 4(||\boldsymbol{\Delta}|| + \alpha)$ "good" stationary point and the stationary point satisfies $|z^Hx| \le \frac{4(||\boldsymbol{\Delta}|| + \alpha)}{\epsilon}$ "bad" stationary point. In the following discussion, we use Lemma 4 to show that the spectral initialization $\text{sgn}(v)$ outperforms all the "bad" fixed points. Due to Lemma 7, Generalized Power Method consistently improve the cost function and thus only converge to "good" stationary points. From Lemma 4, we have

$$\begin{aligned} \text{sgn}(v)^H\boldsymbol{C}\text{sgn}(v) &= |\text{sgn}(v)^Hz|^2 + \text{sgn}(v)^H\boldsymbol{\Delta}\text{sgn}(v) \\ &\ge \left(\epsilon N - \frac{32||\boldsymbol{\Delta}||^2}{\epsilon^3 N}\right)^2 - N||\boldsymbol{\Delta}|| \ge \epsilon^2N^2 - \frac{64||\boldsymbol{\Delta}||^2}{\epsilon^3} - N||\boldsymbol{\Delta}|| \end{aligned} \tag{11}$$

The first inequality is because of Lemma 4, which proved the estimation $\text{sgn}(v)^H\text{sgn}(z) \ge N - \frac{32||\boldsymbol{\Delta}||^2}{\epsilon^2 N}$ and leads to the following results $\text{sgn}(v)^Hz \ge \epsilon N - \frac{32||\boldsymbol{\Delta}||^2}{\epsilon^3 N}$. At the same time, all the bad fixed points x satisfies

$$x^H\boldsymbol{C}x = |x^Hz|^2 + x^H\boldsymbol{\Delta}x \le \frac{64||\boldsymbol{\Delta}||^2}{\epsilon^2} + N||\boldsymbol{\Delta}|| \tag{12}$$

Combine (11), (12) with the assumption $||\boldsymbol{\Delta}|| \le \frac{\epsilon^{3/2}N}{13}$ and $\alpha < ||\boldsymbol{\Delta}||$, we knows that the spectral initialization surpasses all the bad local points. □

For Normalized Generalized Power Method, we can prove a similar version.

Lemma 10. *If x is a fixed point of Normalized Generalized Power Methods (Algorithm 3), at least one of the following holds*

$$|z^Hx| \ge N - \frac{4}{\epsilon}(||\boldsymbol{\Delta}|| + \alpha) \qquad \text{or} \qquad |z^Hx| \le \frac{4}{\epsilon}(||\boldsymbol{\Delta}|| + \alpha).$$

Furthermore, if $||\boldsymbol{\Delta}|| \le \frac{n\epsilon}{13}$ and $\alpha < ||\boldsymbol{\Delta}||$, all the accumulation points x of Algorithm 2 satisfies $|z^Hx| \ge n - 8||\boldsymbol{\Delta}||$.

Proof. If x is a fixed point of Algorithm 3, then x satisfies $||\mathring{\boldsymbol{C}}x||_1 = x^H\mathring{\boldsymbol{C}}x$ (because $|(\mathring{\boldsymbol{C}}x)_i| = \left\langle x_i, (\mathring{\boldsymbol{C}}x)_i \right\rangle$ holds for all i) where $\mathring{\boldsymbol{C}} = \text{sgn}(z)\text{sgn}(z)^H + \text{diag}(\frac{1}{|z|})\boldsymbol{\Delta}\text{diag}(\frac{1}{|z|}) + \alpha\text{diag}(\frac{1}{|z|^2})$. For simplicity, we denote $\mathring{\boldsymbol{\Delta}} = \text{diag}(\frac{1}{|z|})\boldsymbol{\Delta}\text{diag}(\frac{1}{|z|})$ in the following proof.

On one hand

$$x^H\mathring{\boldsymbol{C}}x = |\text{sgn}(z)^Hx|^2 + x^H\mathring{\boldsymbol{\Delta}}x + \alpha\sum_{i=1}^{N}\frac{1}{|z_i|} \le |\text{sgn}(z)^Hx|^2 + \frac{N}{\epsilon}(||\boldsymbol{\Delta}|| + \alpha)$$

On the other hand

$$||\mathring{\boldsymbol{C}}x||_1 = \sum_{i=1}^{N}|(\text{sgn}(z)^Hx)\text{sgn}(z_i) + (\mathring{\boldsymbol{\Delta}}x)_i + \alpha x./|z|| \ge N|\text{sgn}(z)^Hx| - ||\mathring{\boldsymbol{\Delta}}x||_1 - \frac{\alpha N}{\epsilon}$$

At the same time, $||\mathring{\mathbf{\Delta}}x||_1 \leq \sqrt{N}||\mathring{\mathbf{\Delta}}x||_2 \leq N||\mathring{\mathbf{\Delta}}||$. Combine this with the two previous inequality, we get

$$|z^H x|(N - |z^H x|) \leq N||\mathring{\mathbf{\Delta}}|| + \frac{N}{\epsilon}(||\mathbf{\Delta}|| + 2\alpha) \leq \frac{2N}{\epsilon}(||\mathbf{\Delta}|| + \alpha).$$

The above inequality enforces that one of $|z^H x| \geq N - \frac{4}{\epsilon}(||\mathbf{\Delta}|| + \alpha)$ and $|z^H x| \leq \frac{4}{\epsilon}(||\mathbf{\Delta}|| + \alpha)$ holds. We call all the stationary point satisfies $|z^H x| \geq N - \frac{4}{\epsilon}(||\mathbf{\Delta}|| + \alpha)$ good stationary point and the stationary point satisfies $|z^H x| \leq \frac{4}{\epsilon}(||\mathbf{\Delta}|| + \alpha)$ bad stationary point. In the following discussion, we use Lemma 4 to show that the spectral initialization $\text{sgn}(v)$ outperforms all the bad fixed points in terms of the potential function $\mathring{f}(x) = \frac{1}{2}x^H \left(\text{diag}(\frac{1}{|z|})C\text{diag}(\frac{1}{|z|})\right) x$. From Lemma 4, we have

$$\begin{aligned}\mathring{f}(\text{sgn}(v)) &= |\text{sgn}(v)^H z|^2 + \text{sgn}(v)^H \mathring{\mathbf{\Delta}}\text{sgn}(v) \\ &\geq \left(N - \frac{32||\mathbf{\Delta}||^2}{\epsilon^2 N}\right)^2 - \frac{N}{\epsilon}||\mathbf{\Delta}|| \geq N^2 - \frac{64||\mathbf{\Delta}||^2}{\epsilon^2} - \frac{N}{\epsilon}||\mathbf{\Delta}||\end{aligned} \tag{13}$$

The second equality is because $\frac{64||\mathbf{\Delta}||^2}{\epsilon^2 N} \leq ||\text{sgn}(v) - z||^2 \leq 2(N - |z^H \text{sgn}(v)|)$. At the same time, all the bad fixed points x satisfies

$$\mathring{f}(x) = |x^H \text{sgn}(z)|^2 + x^H \mathring{\mathbf{\Delta}} x \leq \frac{64||\mathbf{\Delta}||^2}{\epsilon^2} + \frac{N||\mathbf{\Delta}||}{\epsilon} \tag{14}$$

Combine (13), (14) with the assumption $||\mathbf{\Delta}|| \leq \frac{\epsilon N}{13}$ and $\alpha < ||\mathbf{\Delta}||$, we knows that the spectral initialization surpasses all the bad local points. □

Lemma 1 and Lemma 2 guaranteed the global optimality of the second order stationary point of problem (10). Thus in the next theorem, we verify the Hessian $S(x) = \text{ddiag}(Cxx^H) - C$ the Riemann Hessian $S_{\mathcal{R}}(x) = \text{ddiag}(\mathring{\boldsymbol{C}}xx^H) - \mathring{\boldsymbol{C}}$ is P.S.D over $\mathcal{T}_x\mathbb{T}^n$ at the final stationary point. Then we can conclude the global optimality of the converging point of the Generalized Power Method.

Theorem 6. *Given vector $z \in \mathbb{R}^N$. For matrix $C = zz^\top + \mathbf{\Delta}$, where $\mathbf{\Delta} \in \mathbb{R}^{N\times N}$ is a symmetric perturbation matrix. If $1 - \frac{\sqrt{3}}{2} < \epsilon \leq \min_{i\in[N]} |z_i|$, $||\mathbf{\Delta}|| \leq \frac{\epsilon'}{28}N$ and $||\mathbf{\Delta}||_\infty \leq \frac{\epsilon'}{28}N$, where $\epsilon' = (\epsilon^2 + 2\epsilon - 2)$. When $\alpha \leq ||\mathbf{\Delta}||$, then the GPM converge to the unique global optimum in the quotient space $\mathbb{R}^n/\sim$.*

Proof. From Lemma 7, the Generalized Power Method must converge to a stationary point x. The fixed point of the generalized power methods satisfies $Sx = 0$, which leads to $(\tilde{\boldsymbol{C}}x)_i x_i = |(\tilde{\boldsymbol{C}}x)_i|$ and Lemma 9 guarantees convergence to the good stationary points satisfies $|z^H x| \geq \epsilon N - 4(||\mathbf{\Delta}|| + \alpha)$.

From Lemma 1, we also know that, to prove global optimality of x, it suffices to show that $u^H \boldsymbol{S} u > 0$ holds for all $u \in \mathbb{C}^n$ such that $u \neq 0$ and $u^H x = 0$. This is because

$$\begin{aligned}u^H \boldsymbol{S} u &= \sum_{i=1}^N |u_i|^2 |(C_i x)_i| - u^T \boldsymbol{C} u \\ &= \sum_{i=1}^n |u_i|^2 \left||z^H x| z_i + (\mathbf{\Delta}x)_i\right| - |u^H z|^2 - u^H \mathbf{\Delta} u \\ &\geq \sum_{i=1}^n |u_i|^2 \left(\epsilon|z^H x| - |(\mathbf{\Delta}x)_i|\right) - |u^H(z - x)|^2 - u^H \mathbf{\Delta} u \\ &\geq ||u||^2 \left(\epsilon|z^H x| - ||\mathbf{\Delta}x||_\infty - ||z - x||_2^2 - ||\mathbf{\Delta}||\right) \\ &\geq ||u||^2 \left((2 + \epsilon)(\epsilon N - 4(||\mathbf{\Delta}|| + \alpha)) - 2N - ||\mathbf{\Delta}x||_\infty - ||\mathbf{\Delta}||\right) \\ &\geq ||u||^2((\epsilon^2 + 2\epsilon - 2)N - (9 + 4\epsilon)||\mathbf{\Delta}|| - ||\mathbf{\Delta}||_\infty)\end{aligned} \tag{15}$$

Based on the assumption $||\mathbf{\Delta}|| \leq \frac{\epsilon'}{28}N$ and $||\mathbf{\Delta}z||_\infty \leq \frac{\epsilon'}{28}N$, we know that $u^H \boldsymbol{S} u > 0$. □

Theorem 7. *Given vector $z \in \mathbb{R}^N$ and there exists a constant $\epsilon > 0$ such that $\epsilon \leq \min_{i\in[N]} |z_i|$. For matrix $C = zz^\top + \mathbf{\Delta}$, where $\mathbf{\Delta} \in \mathbb{R}^{N\times N}$ is a symmetric perturbation matrix. If $||\mathbf{\Delta}|| \leq \frac{\epsilon}{28}N$ and $||\mathbf{\Delta}||_\infty \leq \frac{\epsilon}{28}N$, when $\alpha \leq ||\mathbf{\Delta}||$, then the normalized GPM converge to the unique global optimum in the quotient space $\mathbb{R}^n/\sim$.*

Proof. From Lemma 8, the Normalized Generalized Power Methods must converge to a stationary point x. The first order condition of the stationary point is $||\mathring{\boldsymbol{C}}x||_1 = x^H\mathring{\boldsymbol{C}}x$ and Lemma 10 guarantees convergence to the good stationary points satisfies $|z^Hx| \geq N - \frac{4}{\epsilon}(||\boldsymbol{\Delta}|| + \alpha)$.

Observe that $S_{\mathcal{R}}(x)x = 0$, according to Lemma 2, the only thing we need to prove global optimality of the converged stationary point x is to verify the $u^H S_{\mathcal{R}}(x)u > 0$ for all $u^Hx = 0$.

$$\begin{aligned}
u^H\boldsymbol{S}u &= \sum_{i=1}^{N}|u_i|^2|(\mathring{\boldsymbol{C}}_ix)_i| - u^T\mathring{\boldsymbol{C}}u \\
&= \sum_{i=1}^{n}|u_i|^2\left||\text{sgn}(z)^Hx|\text{sgn}(z_i) + (\mathring{\boldsymbol{\Delta}}x)_i\right| - |u^H\text{sgn}(z)|^2 - u^H\mathring{\boldsymbol{\Delta}}u \\
&\geq \sum_{i=1}^{n}|u_i|^2\left(|\text{sgn}(z)^Hx| - |(\mathring{\boldsymbol{\Delta}}x)_i|\right) - |u^H(\text{sgn}(z) - x)|^2 - u^H\boldsymbol{\Delta}u - \alpha||u||_2^2 \\
&\geq ||u||^2\left(|\text{sgn}(z)^Hx| - ||\mathring{\boldsymbol{\Delta}}x||_\infty - ||\text{sgn}(z) - x||_2^2 - ||\mathring{\boldsymbol{\Delta}}||\right) \\
&\geq ||u||^2\left(N - \frac{12}{\epsilon}(||\boldsymbol{\Delta}|| + \alpha) - \frac{1}{\epsilon}||\boldsymbol{\Delta}||_\infty - \frac{1}{\epsilon}||\boldsymbol{\Delta}||\right)
\end{aligned} \tag{16}$$

Based on the assumption $||\boldsymbol{\Delta}|| \leq \frac{\epsilon}{28}N$ and $||\boldsymbol{\Delta}z||_\infty \leq \frac{\epsilon}{28}N$, we know that $u^H\boldsymbol{S}u > 0$. □

8.2.3 Linear Rate Convergence

In this section, following Liu et al. [2017], we provide the proof of linear rate convergence of the normalized Generalized Power Method on our problem. With out loss of generality, we assume $1 = \arg\min_{\theta\in\{1,-1\}}||\theta y^k - z||_2$ for all $k \in \mathbb{N}$, where $\{y^k\}_{k\in\mathbb{N}}$ is the iterates generated by the normalized generalized power method.

Theorem 8 (Estimation Bound). *Suppose that $||\boldsymbol{\Delta}|| \leq \frac{N\epsilon}{16}$ and $\alpha < \frac{N\epsilon}{6}$. Then the iterates $\{y^k\}_{k\in\mathbb{N}}$ generated by the normalized generalized power method satisfies*

$$||y^{k+1} - \text{sgn}(z)|| \leq \mu^{k+1}||y^0 - \text{sgn}(z)|| + \frac{\nu}{1-\mu}\frac{8||\boldsymbol{\Delta}||}{\epsilon\sqrt{N}}$$

for all $k \in \mathbb{N}$, where

$$\mu = \frac{16(\alpha + ||\boldsymbol{\Delta}||)}{(7N\epsilon - 8\alpha)} < 1, \nu = \frac{2N}{7N - 8\frac{\alpha}{\epsilon}}.$$

Proof. This Theorem is a direct adaptation of [Liu et al., 2017, Theorem 3.1]. □

Based on the previous estimation bound. We can build the local error bounds to guarantees global convergence. The local error bounds provide an estimation of the distance between any points in sgn(z)'s neighborhood and the global optima of the original optimization problem. To do this, we first define two mappings $\Sigma : \mathbb{T}^n \to \mathbb{H}^n$ and $\rho : \mathbb{T}^n \to \mathbb{R}+$ as

$$\Sigma(z) = \text{diag}(|\mathring{\boldsymbol{C}}z|) - \mathring{\boldsymbol{C}}, \quad \rho(z) = ||\boldsymbol{\Sigma}(z)z||_2.$$

Then we can have the following results

Lemma 11. *We denote z^* the global optimum of problem (10) and $\{y^{(k)}\}_{k\in\mathbb{N}}$ the iterates generated by the Normalized Generalized Power Method. If $\alpha \leq ||\boldsymbol{\Delta}|| \leq \frac{\epsilon}{216}N$ and $||\boldsymbol{\Delta}||_\infty \leq \frac{\epsilon}{12}N$, then we have*

- *(Local Error Bound) $||y - z^*|| \leq \frac{N}{4}\rho(y)$ holds for all*
- *$\rho(y^k) \leq a||y^{k+1} - y^k||_2$ holds for some constant a.*

Proof. **Proof of Local Error Bound** To prove the local error bound, we make the following decomposition $||\mathbf{\Sigma}(y)y|| \geq ||\mathbf{\Sigma}(z^*)y|| - ||(\mathbf{\Sigma}(y) - \mathbf{\Sigma}(z^*))y||$. We first build the lower bound of $||(\mathbf{\Sigma}(y) - \mathbf{\Sigma}(z^*))y||$ following [Liu et al., 2017, Proposition 4.2]

$$\begin{aligned}||(\mathbf{\Sigma}(y) - \mathbf{\Sigma}(z^*))y|| &= |||\mathring{\boldsymbol{C}}y| - |\mathring{\boldsymbol{C}}z^*|||_2 \leq ||\mathring{\boldsymbol{C}}(y - \hat{z})|| \\ &\leq \sqrt{N}|z^H(y - z^*)| + \frac{\alpha + ||\mathbf{\Delta}||}{\epsilon}||y - z^*|| \\ &\leq \sqrt{N}|(z^H - z^*)^H(y - z^*)| + \sqrt{N}|(z^*)^H(y - z^*)| + \frac{\alpha + ||\mathbf{\Delta}||}{\epsilon}||y - z^*|| \\ &\leq \frac{\alpha + 5||\mathbf{\Delta}||}{\epsilon}||y - z^*|| + \frac{1}{2}||y - z^*||^2\end{aligned} \tag{17}$$

Similar to Theorem 7, we can then lower bound $||\Sigma(z^*)y|| = ||\Sigma(z^*)\hat{y}||$ where $\hat{y} = (I - \frac{1}{n}z^*(z^*)^H)(y - z^*)$ is the projection of $y - z^*$ onto the orthogonal complement of span$(\hat{z})$. At the same time

$$||\hat{u}|| \geq ||y - z^*|| - \left|\left|\frac{1}{n}z^*(z^*)^H(y - z^*)\right|\right| = ||y - z^*|| - \frac{||y - z^*||^2}{2\sqrt{N}}$$

where the last equality is because $||y - z^*||^2 = 2(N - |y^H z^*|)$. Similar to Theorem 7, we have

$$\begin{aligned}||\hat{y}||||\Sigma(z^*)\hat{y}|| \geq \hat{y}^H\Sigma(z^*)\hat{y} &= \hat{y}^H(\text{diag}(|\mathring{\boldsymbol{C}}z|) - \mathring{\boldsymbol{C}})\hat{y} \\ &= \sum_{i=1}^{n}|\hat{y}_i|^2\left||\text{sgn}(z)^H x|\text{sgn}(z_i) + (\mathring{\mathbf{\Delta}}x)_i\right| - |\hat{y}^H\text{sgn}(z)|^2 - \hat{y}^H\mathring{\mathbf{\Delta}}\hat{y} \\ &\geq ||\hat{y}||^2\left(N - \frac{12}{\epsilon}(||\mathbf{\Delta}|| + \alpha) - \frac{1}{\epsilon}||\mathbf{\Delta}||_\infty - \frac{1}{\epsilon}||\mathbf{\Delta}||\right).\end{aligned} \tag{18}$$

Thus

$$||\Sigma(z^*)\hat{y}|| \geq \left(||y - z^*|| - \frac{||y - z^*||^2}{2\sqrt{N}}\right)\left(N - \frac{12}{\epsilon}(||\mathbf{\Delta}|| + \alpha) - \frac{1}{\epsilon}||\mathbf{\Delta}||_\infty - \frac{1}{\epsilon}||\mathbf{\Delta}||\right).$$

At the same time $||y - z^*||^2 \leq ||y - z^*||(||y - z|| + ||z - z^*||) \leq \left(\frac{\sqrt{N}}{2} + \frac{4||\mathbf{\Delta}||}{\sqrt{N}}\right)||y - z^*||$. Combining all the results we get and finally we have

$$\rho(z) \geq \left[\frac{N}{2} - \frac{18(||\mathbf{\Delta}|| + \alpha)}{\epsilon} - \frac{||\mathbf{\Delta}||}{\epsilon}\right]||y - z^*||. \tag{19}$$

Based on the assumptions $\alpha \leq ||\mathbf{\Delta}|| \leq \frac{\epsilon}{216}N$ and $||\mathbf{\Delta}||_\infty \leq \frac{\epsilon}{12}N$, we know that $\rho(z) \geq \frac{N}{4}d(z, \hat{z})$ **Proof of** $\rho(y^k) \leq a||y^{k+1} - y^k||_2$ By definition of y^{k+1}, $\rho(y^k) = ||\text{diag}(|\mathring{\boldsymbol{C}}y^k|)(y^{k+1} - y^k)|| \leq ||\text{diag}(|\mathring{\boldsymbol{C}}y^k|)||_\infty||y^{k+1} - y^k||$. At the same time, we have

$$\begin{aligned}||\text{diag}(|\mathring{\boldsymbol{C}}y^k|)||_\infty &\leq ||zz^H y^k||_\infty + \frac{\alpha + ||\mathbf{\Delta}||_\infty}{\epsilon} \\ &= |z^H y^k| + \frac{\alpha + ||\mathbf{\Delta}||_\infty}{\epsilon} \leq 2N.\end{aligned} \tag{20}$$

This leads to the estimate $\rho(y^k) \leq 2N||y^{k+1} - y^k||_2$

□

Theorem 9. *We make the same assumption as Lemma 11. We further assumes $\mathring{\boldsymbol{C}} \succeq a_0\boldsymbol{I}$ for some constant $a' > 0$, then the normalized generalized power method linearly converge to the global optimum z^*.*

Remark 6. *In Liu et al. [2017], the data generating process doesn't ensures the matrix C is P.S.D. Thus Liu et al. [2017] should apply a lower bound on α to ensure $\mathring{\boldsymbol{C}}$ is P.S.D. In our case, the matrix is the covariance matrix of a noisy dataset. Thus it is nature have P.S.D. $\mathring{\boldsymbol{C}}$.*

Proof. For $\mathring{\boldsymbol{C}} \succeq a_0 I$, it's obvious to have sufficient ascent $\mathring{f}(y^{k+1}) - \mathring{f}(y^k) \geq a_0||y^{k+1} - y^k||_2^2$ holds for every iteration ([Boumal, 2016, Lemma 8],[Liu et al., 2017, Proposition 4.3(a)]). Thus $f(y^{k+1}) - f(y^k) \geq \epsilon a_0||y^{k+1} - y^k||_2^2$. Before we

present the final linear convergence proof, we first prove that $f(z^*) - f(y^k) \leq a_1 ||y^k - z^*||^2$. This is because

$$\begin{aligned} f(z^*) - f(y^k) &\leq \frac{1}{\epsilon}(\mathring{f}(z^*) - \mathring{f}(y^k)) \\ &= (y^k)^H \left(\text{diag}(|\mathring{\boldsymbol{C}} z^*|) - \mathring{\boldsymbol{C}}\right) y^k \\ &= (y^k - z^*)^H \left(\text{diag}(|\mathring{\boldsymbol{C}} z^*|) - \mathring{\boldsymbol{C}}\right) (y^k - z^*) \\ &\leq (||\mathring{\boldsymbol{C}}|| + ||\mathring{\boldsymbol{C}}||_\infty) ||y^k - z^*||^2. \end{aligned} \tag{21}$$

Now we are equipped with all the inequality needed to provide a global convergence proof. According to [Liu et al., 2017, Proof of Theorem 4.1], we knows that the normalized generalized power methods convergence to the global optimum linearly. □

8.2.4 Generative Models

In this section, we'll discuss how the random data sampled form the linear fixed effect model (also referred to an "interactive fixed-effect model") satisfies the discordant assumptions we used to prove the global optimization results. We assume the outcomes are generated via the following linear factor model

$$Y_{jt} = \boldsymbol{\Delta}_t + D_{jt}\tau + \theta_t^T \mu_j + e_{jt}, \qquad \mathbb{E}[e_{jt}|\boldsymbol{\Delta}_t, \mu_j, D_{jt}] = 0, \text{Var}[e_{jt}|\boldsymbol{\Delta}_t, \mu_j, D_{jt}] = \sigma$$

where $\boldsymbol{\Delta}_t$ is the time fixed effect, μ_j is the unobserved common factors and θ_t is a vector of unknown factor loadings. ϵ_{jt} is the unobserved idiosyncratic noise. τ is the treatment effect we aim to estimate and D_{jt} is the 0-1 variable according to the treatment assignment to unit j at time t. In specific, in the pre-treatment period, $D_{jt} = 0$ for all $\forall j \in [N], t \in [T]$. Thus the outcome matrix $Y \in \mathbb{R}^{N \times T}$ can be written in the following compact matrix form

$$Y = \underbrace{\begin{bmatrix} \mu_1^\top & 1 \\ \mu_2^\top & 1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \mu_N^\top & 1 \end{bmatrix}}_{\mu} \underbrace{\begin{bmatrix} \theta_1 & \cdots & \theta_t \\ \boldsymbol{\Delta}_1 & \cdots & \boldsymbol{\Delta}_t \end{bmatrix}}_{\theta} + W$$

where W is a matrix whose entries are i.i.d. standard normal random variables denote the measurement noise. We consider the time factor is sampled from a underlying distribution $\begin{pmatrix} \theta_i \\ \boldsymbol{\Delta}_i \end{pmatrix} \sim p(\tilde{\theta}, \tilde{\Sigma})$ and $\Sigma_\theta \triangleq \tilde{\theta}\tilde{\theta}^\top + \tilde{\Sigma} = \mathbb{E} \begin{pmatrix} \theta_i \\ \boldsymbol{\Delta}_i \end{pmatrix} \begin{pmatrix} \theta_i \\ \boldsymbol{\Delta}_i \end{pmatrix}^\top$. Then we knows that $\mathbb{E}\boldsymbol{Y}\boldsymbol{Y}^\top = \mu \Sigma_\theta \mu^\top + \sigma I_N$. We first assume that Σ_θ is a non-degenerate covariance matrix.

Assumption 4. *Σ_θ is positive semi-definite.*

Assumption 2 means that the matrix $\Sigma \triangleq \mu \Sigma_\theta \mu^\top$ is rank $n-1$. We assume $v = (w_i D_i)_{i=1}^n$ to be vector in the null space of Σ, where $(w_i, D_i)_{i=1}^n$ is the only realizable experiment profile in Assumption 2. To verify that the data generating processing satisfies the assumptions we made for global convergence. We further made the following assumptions to the regularity of the problem.

Assumption 5 (Regularity of the Problem)**.** *We further assume the following regularity properties of the covariance matrix and random sample Y_t*

- $||\boldsymbol{\Sigma}^\dagger|| \leq C_1$ *holds for some constant* C_1.
- $||\boldsymbol{\Sigma}^\dagger||_\infty \leq O(N^{c_1})$ *holds for some constant* $c_1 \geq 0$.
- $||Y_t Y_t^\top|| \leq C_2$ *holds almost surely holds for some constant* C_2.
- $||Y_t Y_t^\top||_\infty \leq O(N^{c_2})$ *holds almost surely for some constant* $c_2 \geq 0$.

Bound $||\sigma N(\boldsymbol{Y}\boldsymbol{Y}^\top + \sigma I)^{-1} - uu^\top||$ In the following paragraph, we bound the error between the iteration matrix with the rank one ground truth in ℓ_2 operator norm. To do this, we make the following decomposition

$$\begin{aligned} ||\sigma N(\boldsymbol{Y}\boldsymbol{Y}^\top + \sigma I)^{-1} - uu^\top|| &\leq \sigma N||(\boldsymbol{Y}\boldsymbol{Y}^\top + \sigma I)^{-1} - (\boldsymbol{\Sigma} + \sigma I)^{-1}|| + ||\sigma N(\boldsymbol{\Sigma} + \sigma I)^{-1} - uu^\top|| \\ &\leq \sigma N||(\boldsymbol{\Sigma} + \sigma I)^{-1}||||\boldsymbol{Y}\boldsymbol{Y}^\top - \boldsymbol{\Sigma}|| + ||\sigma N(\boldsymbol{\Sigma} + \sigma I)^{-1} - uu^\top|| \end{aligned} \tag{22}$$

We first bound $||\sigma N(\boldsymbol{\Sigma} + \sigma I)^{-1} - uu^\top||$. To bound this term, we use the geometric series expansion $\frac{\lambda}{\lambda + X} = \sum_{j=0}^{\infty}(-1)^j \left(\frac{\lambda}{X}\right)^{j+1}$. If $\sigma||\boldsymbol{\Sigma}^\dagger|| < 1$, then

$$||\sigma N(\boldsymbol{\Sigma} + \sigma I)^{-1} - uu^\top|| \leq N\sum_{j=0}^{\infty}\sigma^{j+1}||\boldsymbol{\Sigma}^\dagger||^{j+1} = \frac{N\sigma||\boldsymbol{\Sigma}^\dagger||}{1 - \sigma||\boldsymbol{\Sigma}^\dagger||} \tag{23}$$

To bound $||\sigma N(\boldsymbol{\Sigma} + \sigma I)^{-1}||||\boldsymbol{Y}\boldsymbol{Y}^\top - \boldsymbol{\Sigma}||$, we first use the matrix Bernstein inequality Tropp [2015], Tao [2011] to bound $||\boldsymbol{Y}\boldsymbol{Y}^\top - \boldsymbol{\Sigma}||$. We know

$$||\boldsymbol{Y}\boldsymbol{Y}^\top - \boldsymbol{\Sigma}|| \leq \sqrt{\frac{C_2^2\log(\boldsymbol{\Delta})}{T}} + \frac{2C_2\log(\boldsymbol{\Delta})}{T}$$

with high probability $1 - \boldsymbol{\Delta}$. At the same time, we have

$$||\sigma N(\boldsymbol{\Sigma} + \sigma I)^{-1}|| \leq ||\sigma N(\boldsymbol{\Sigma} + \sigma I)^{-1} - uu^\top|| + ||uu^\top|| \leq \frac{N}{(1 - \sigma||\boldsymbol{\Sigma}^\dagger||_\infty)}.$$

Finally, we achieve

$$||\sigma N(\boldsymbol{Y}\boldsymbol{Y}^\top + \sigma I)^{-1} - uu^\top|| \leq \frac{N\sigma||\boldsymbol{\Sigma}^\dagger||}{1 - \sigma||\boldsymbol{\Sigma}^\dagger||} + \frac{N}{(1 - \sigma||\boldsymbol{\Sigma}^\dagger||_\infty)}\sqrt{\frac{C_2^2\log(\boldsymbol{\Delta})}{T}}$$

holds with high probability $1 - \boldsymbol{\Delta}$.

Bound $||\sigma N(\boldsymbol{Y}\boldsymbol{Y}^\top + \sigma I)^{-1} - uu^\top||_\infty$ In the following paragraph, we bound the error between the iteration matrix with the rank one ground truth in ℓ_∞ operator norm.

$$\begin{aligned} ||\sigma N(\boldsymbol{Y}\boldsymbol{Y}^\top + \sigma I)^{-1} - uu^\top||_\infty &\leq \sigma N||(\boldsymbol{Y}\boldsymbol{Y}^\top + \sigma I)^{-1} - (\boldsymbol{\Sigma} + \sigma I)^{-1}||_\infty + ||\sigma N(\boldsymbol{\Sigma} + \sigma I)^{-1} - uu^\top||_\infty \\ &\leq \sigma N||(\boldsymbol{\Sigma} + \sigma I)^{-1}||||\boldsymbol{Y}\boldsymbol{Y}^\top - \boldsymbol{\Sigma}||_\infty + ||\sigma N(\boldsymbol{\Sigma} + \sigma I)^{-1} - uu^\top||_\infty \end{aligned} \tag{24}$$

We first bound $||\sigma N(\boldsymbol{\Sigma} + \sigma I)^{-1} - uu^\top||_\infty$. To bound this term, we use the geometric series expansion $\frac{\lambda}{\lambda + X} = \sum_{j=0}^{\infty}(-1)^j \left(\frac{\lambda}{X}\right)^{j+1}$. If $\sigma||\boldsymbol{\Sigma}^\dagger||_\infty < 1$, then

$$||\sigma N(\boldsymbol{\Sigma} + \sigma I)^{-1} - uu^\top||_\infty \leq N\sum_{j=0}^{\infty}\sigma^{j+1}||\boldsymbol{\Sigma}^\dagger||_\infty^{j+1} = \frac{N\sigma||\boldsymbol{\Sigma}^\dagger||_\infty}{1 - \sigma||\boldsymbol{\Sigma}^\dagger||_\infty} \tag{25}$$

To bound $||\sigma N(\boldsymbol{\Sigma} + \sigma I)^{-1}||||\boldsymbol{Y}\boldsymbol{Y}^\top - \boldsymbol{\Sigma}||_\infty$, we first use the matrix Bernstein inequality Tropp [2015], Tao [2011] to bound $||\boldsymbol{Y}\boldsymbol{Y}^\top - \boldsymbol{\Sigma}||_\infty$. We know

$$||\boldsymbol{Y}\boldsymbol{Y}^\top - \boldsymbol{\Sigma}||_\infty \leq \sqrt{\frac{N^{2c_2}\log(\boldsymbol{\Delta})}{T}} + \frac{2N^{c_2}\log(\boldsymbol{\Delta})}{T}$$

with high probability $1 - \boldsymbol{\Delta}$. At the same time, we have

$$||\sigma N(\boldsymbol{\Sigma} + \sigma \boldsymbol{I})^{-1}|| \leq ||\sigma N(\boldsymbol{\Sigma} + \sigma \boldsymbol{I})^{-1} - uu^\top||_\infty + ||uu^\top||_\infty \leq \frac{N}{\epsilon^2(1 - \sigma||\boldsymbol{\Sigma}^\dagger||_\infty)}.$$

We plug in all the bounds and finally get

$$||\sigma N(\boldsymbol{Y}\boldsymbol{Y}^\top + \sigma \boldsymbol{I})^{-1} - uu^\top||_\infty \leq \frac{N\sigma||\boldsymbol{\Sigma}^\dagger||_\infty}{1 - \sigma||\boldsymbol{\Sigma}^\dagger||_\infty} + \frac{N}{\epsilon^2(1 - \sigma||\boldsymbol{\Sigma}^\dagger||_\infty)}\sqrt{\frac{N^{2c_2}\log(\boldsymbol{\Delta})}{T}}$$

with high probability.

From the discussion in Appendix 8.2.3, if we can bound both $||\sigma N(\boldsymbol{Y}\boldsymbol{Y}^\top + \sigma I)^{-1} - uu^\top||$ and $||\sigma N(\boldsymbol{Y}\boldsymbol{Y}^\top + \sigma I)^{-1} - uu^\top||_\infty$ as $O(\epsilon N)$, then we can have global convergence results. It's easy to check that if we select $\sigma \leq \Omega(\epsilon N^{-c_1}), T \geq \Omega(\epsilon^6 N^{2c_2})$, then the assumptions for global convergence holds.

Corollary 1. *If $c_1 = 0$,* i.e. *there exists some constant C_1 such that $||\Sigma^\dagger||_\infty \leq C_1$, then the noise level $\sigma \leq \Omega(\epsilon)$ and $T > \Omega(\epsilon^6 N^{2c_2})$ ensures global convergence of NormSPCD algorithm.*

Algorithm 4 Empirical Implementation of SPCD

Require: Pre-treatment Observations $Y \in \mathbb{R}^{T\times N}$

Set initial treatment assignment guess through $y^0 = \text{sgn}(v)$, where v is the smallest eigenvector of matrix $(\boldsymbol{Y}\boldsymbol{Y}^\top + \alpha I + \lambda \mathbb{1}\mathbb{1}^\top)$, where α, λ are two pre-defined hyper-parameter.

▷ Spectral Initialization

while Converged **do**

Select one of the following two boxes to iterate

For SPCD, update the design via ▷ **Generalized power methods**

$$y^{t+1} = \text{sgn}\left[\left((\boldsymbol{Y}\boldsymbol{Y}^\top + \alpha \boldsymbol{I} + \lambda \mathbb{1}\mathbb{1}^\top)^{-1} + \beta \boldsymbol{I}\right) y^t\right], \tag{26}$$

where β is a pre-defined hyper-parameter.

For NormSPCD, update the design via ▷ **Normalize the inverse covariance matrix**

$$y^{t+1} = \text{sgn}\left[\left[(\boldsymbol{Y}\boldsymbol{Y}^\top + \alpha \boldsymbol{I} + \lambda \mathbb{1}\mathbb{1}^\top)^{-1} + \beta \boldsymbol{I}\right] (y^t/d)\right], \tag{27}$$

where $d = \sqrt{\text{diag}((\boldsymbol{Y}\boldsymbol{Y}^\top + \alpha \boldsymbol{I} + \lambda \mathbb{1}\mathbb{1}^\top)^{-1})}$ and $/$ denotes element-wise divide.

end while

Once obtained the optimal design y^*, one can select the design weight w via

$$w = \frac{2(\boldsymbol{Y}\boldsymbol{Y}^\top + \alpha \boldsymbol{I} + \lambda \mathbb{1}\mathbb{1}^\top)^{-1} y^*}{\|(\boldsymbol{Y}\boldsymbol{Y}^\top + \alpha \boldsymbol{I} + \lambda \mathbb{1}\mathbb{1}^\top)^{-1} y^*\|_1} \tag{28}$$

▷ The optimality condition ensures $\text{sgn}(w) = y$.

Treat Unit i if $y(i) = -\text{sgn}\left(\sum_{i=1}^N y(i)\right)$ and run the experiment.

Estimate the treatment effect via

$$\hat{\tau} = \sum_{t=1}^{S}\left(\sum_{i=1}^{N} w(i) Y_{i,T+t}\right)$$
