Gaussianized Design Optimization for Covariate Balance in Randomized Experiments

WENXUAN GUO, University of Chicago, USA
TENGYUAN LIANG, University of Chicago, USA
PANOS TOULIS, University of Chicago, USA

Achieving covariate balance in randomized experiments enhances the precision of treatment effect estimation. However, existing methods often require heuristic adjustments based on domain knowledge and are primarily developed for binary treatments. This paper presents Gaussianized Design Optimization, a novel framework for optimally balancing covariates in experimental design. The core idea is to gaussianize the treatment assignments: we model treatments as transformations of random variables drawn from a multivariate Gaussian distribution, converting the design problem into a nonlinear continuous optimization over Gaussian covariance matrices. Compared to existing methods, our approach offers significant flexibility in optimizing covariate balance across a diverse range of designs and covariate types. Adapting the Burer-Monteiro approach for solving semidefinite programs, we introduce first-order local algorithms for optimizing covariate balance, improving upon several widely used designs. Furthermore, we develop inferential procedures for constructing design-based confidence intervals under Gaussianization and extend the framework to accommodate continuous treatments. Simulations demonstrate the effectiveness of Gaussianization in multiple practical scenarios.

Additional Key Words and Phrases: Continuous Treatments, Covariate Balance, Mehler's Formula, Optimal Experimental Design

1 Introduction

Randomized experiments are considered the gold standard for causal inference in the study of treatment effects [Imbens and Rubin, 2015]. Common design choices include completely randomized experiments and independent Bernoulli randomization, which treat experimental units equally and allow for valid estimation of a wide variety of causal quantities. In many real-world experiments, incorporating covariates can enhance balance and improve the precision of treatment effect estimation. Examples of this include matched-pair designs [Bai, 2022, Fisher, 1935] and rerandomization methods [Li et al., 2020, Morgan and Rubin, 2012]. Despite their widespread use, several important design optimization questions related to general covariates and treatments remain underexplored.

By leveraging concepts from continuous optimization and Gaussian processes, we make progress toward addressing two design optimization questions in this paper: (i) How can covariate balance be numerically optimized in experimental designs with general covariates? (ii) How can covariates be balanced with multiple treatment arms, or more generally, in settings involving a continuum of treatment arms?

We propose *Gaussianized Design Optimization*, a framework for experimental design that transforms the design problem into an optimization problem in an embedded Gaussian space. To illustrate the basic idea, suppose we have n experimental units, each receiving a treatment taking values in a discrete space, say, $D_i \in \mathbb{D}$ for $i = 1, \dots, n$. Then, Gaussianization refers to the action of modeling treatments $\{D_i\}_{i=1}^n$ as random variables derived from Gaussian vectors,

$$D_i = g(T_i)\ , \quad T \coloneqq (T_1, \dots, T_n) \sim \mathcal{N}(0, \Sigma)\ .$$

Here $g : \mathbb{R} \to \mathbb{D}$ is a pre-specified function that maps the Gaussian variables T_i to the treatment space, and Σ is a design matrix from the correlation elliptope,

$$\mathcal{E} = \{\ \Sigma \in \mathbb{R}^{n\times n} \mid \Sigma \succeq 0, \Sigma_{ii} = 1\ \}\ . \tag{1}$$

Gaussianization thus transforms the design problem on $\{D_i\}_{i=1}^n$ to a design problem on $\{T_i\}_{i=1}^n$, motivating design optimization in the embedded Gaussian space. Given pre-treatment covariates

$X \in \mathbb{R}^{n\times d}$, at a high level, we propose to solve

$$\min_{\Sigma\in\mathcal{E}} \|X^\top f(\Sigma)X\|_{\text{norm}}\,, \quad \text{norm} \in \{\text{nuc}, \text{op}\}\,. \tag{2}$$

Here, f is a function with an analytical expression applied elementwise, defined later in Sections 1.1 and 3, that controls the aspects of the design that are important for covariate balance. We use nuc and op to abbreviate the nuclear norm and operator norm, respectively. The objective $\|X^\top f(\Sigma)X\|_{\text{norm}}$ serves as a surrogate metric to optimize the covariate balance, which we explain in Section 1.1.

From an optimization perspective, Equation (2) is a nonlinear optimization problem over the correlation elliptope, and we propose a first-order local algorithm to iteratively update the design Σ. Due to the complex, non-convex nature of the covariate balance objective, our algorithm is only guaranteed to find local optimizers near the initial point. This local optimality and the computational barrier of design optimization are further discussed in Section 1.1.

Gaussianization transforms the design problem into an optimization task, providing a flexible framework for optimizing covariate balance. Importantly, this approach applies directly to any number of treatment arms and any type of covariates. In contrast, most existing research on optimal design focuses on binary treatments [Bai, 2023, Harshaw et al., 2019, Li and Ding, 2020], and certain optimality criteria require additional knowledge about the outcome-generating model [Bai, 2022].

In certain experiments, the treatment variable is inherently continuous (e.g., a medication dosage), making it insufficient to confine the design to a small number of discrete arms. To address this limitation, we further extend the discrete treatment setting by allowing $\mathbb{D} = \mathbb{R}$ and propose Gaussian designs, which directly assign $T = (T_1, \ldots, T_n) \sim \mathcal{N}(0, \Sigma)$ as actual treatments. As shown in Section 5, this approach offers two main advantages. First, it allows the exploration of structural properties of potential outcome functions, including monotonicity and convexity. Second, it enables covariate balance optimization akin to the discrete setup. Thus, Gaussian designs harness the flexibility of Gaussianization and contribute to the growing literature on continuous treatment effects.

In Section 6, we investigate design-based inference under Gaussianization, where the outcome-generating model is fixed, and all randomness arises from the Gaussian treatments $\{T_i\}_{i=1}^n$. Under a local perturbation condition, we establish asymptotic normality for the proposed estimator and present valid inferential procedures. Collectively, we establish a comprehensive framework that integrates design optimization, estimation, and inference under Gaussianization.

1.1 An Example: Gaussianization with Three Treatment Arms

To contextualize the idea, we first walk through Gaussianized design optimization with a simple three-treatment example, i.e., $\mathbb{D} = \{1, 2, 3\}$, supplemented with a numerical simulation. Following the standard potential outcome framework [Neyman, 1923], we define $Y_i(k)$ as the potential outcome for unit i under treatment k for $k = 1, 2, 3$. The observed outcome for unit i is then defined as $Y_i = \sum_{k=1}^3 \mathbb{I}\{D_i = k\}Y_i(k)$ and $D = (D_1, \ldots, D_n)$ is the treatment vector. Let $X \in \mathbb{R}^{n\times d}$ be the covariate matrix, and $X_i \in \mathbb{R}^d$ be unit i's covariates.

In this three treatment arms setup, the Gaussianized design optimization framework breaks down to the procedure below. Technical details will be provided in Sections 3 and 4.

Procedure 1 (High-Level Procedure of Gaussianized Design Optimization).

(1) Specify the estimand: Here, we focus on the average treatment effect of all treatment arms

$$\tau = \frac{1}{3}\sum_{k=1}^3 \tau_k\,, \quad \tau_k = \frac{1}{n}\sum_{i=1}^n Y_i(k)\,.$$

We use a Hortivz-Thompson estimator $\widehat{\tau}$ to unbiasedly estimate this quantity.

(2) Derive measures of covariate balance: We propose the following covariate balance measures

$$\sum_{k=1}^{3} \|X^\top \mathrm{Cov}_k(D) X\|_{\mathrm{norm}} , \quad \mathrm{norm} \in \{\mathrm{op}, \mathrm{nuc}\} ,$$

where $\mathrm{Cov}_k(D)$ is the covariance matrix of $(\mathbb{I}\{D_1 = k\}, \ldots, \mathbb{I}\{D_n = k\})$. The measures in the operator and nuclear norm capture the worst-case and average-case mean squared error (MSE) of $\widehat{\tau}$, respectively, as explained in Section 3.

(3) Gaussianizaton: We model treatments by $D_i = g(T_i)$ and derive that $\mathrm{Cov}_k(D) = f_k(\Sigma)$, $k = 1, 2, 3$ with analytical expressions of f_k. These known functions f_k are explicitly given in Proposition 1. The act of Gaussianization translates covariate balance measures to an analytical function on the Gaussian covariance Σ, as follows:

$$\sum_{k=1}^{3} \|X^\top \mathrm{Cov}_k(D) X\|_{\mathrm{norm}} = \sum_{k=1}^{3} \|X^\top f_k(\Sigma) X\|_{\mathrm{norm}} , \quad \mathrm{norm} \in \{\mathrm{op}, \mathrm{nuc}\} .$$

(4) Solve the design optimization: We apply a first-order algorithm (projected gradient descent on the Gaussianized space) in Section 4 to solve

$$\min_{\Sigma \in \mathcal{E}} \sum_{k=1}^{3} \|X^\top f_k(\Sigma) X\|_{\mathrm{norm}} . \tag{3}$$

This returns a locally optimal Gaussian covariance matrix Σ^*.

(5) Assign treatments. Generate treatments through $D_i = g(T_i)$, $T \sim \mathcal{N}(0, \Sigma^*)$.

Optimization and sampling benefits. Procedure 1 applies to general experimental setups with $\mathbb{D} = \{1, \ldots, K\}$, where K is the total number of treatment arms (Section 3). Given covariate balance measures of the form $\sum_{k=1}^{K} \|X^\top \mathrm{Cov}_k(D) X\|_{\mathrm{norm}}$ in Step 2, one would naturally search for a valid design with the optimal covariate balance measure. However, it is unclear how to directly optimize the design for D. First, optimization over the treatment covariance $\mathrm{Cov}_k(D)$ is computationally challenging: for binary treatment assignments ($K = 2$), we need to solve

$$\min_{\mathrm{Cov}_1(D)} \|X^\top \mathrm{Cov}_1(D) X\|_{\mathrm{nuc}} = \min_{\mathrm{Cov}_1(D)} \mathrm{tr}\big(XX^\top \mathrm{Cov}_1(D)\big) \Leftrightarrow \min_{C \in C} \mathrm{tr}\big(XX^\top C\big) .$$

The feasible set of $\mathrm{Cov}_1(D)$ is affinely isomorphic to the cut polytope C [Huber and Maric, 2017], and the optimization problem is thus equivalent to the Max-Cut problem [Barahona and Mahjoub, 1986], which is NP-hard. Second, even if one somehow obtains an approximate solution for $\mathrm{Cov}_k(D)$, it is still unclear how to sample discrete treatment assignments $\{D_i\}_{i=1}^n$ to achieve the desired covariance matrices $\mathrm{Cov}_k(D)$, $k = 1, \ldots, K$.

Gaussianization mitigates these computational and sampling difficulties. Based on the discussion above, Gaussianization transforms the design problem into a nonlinear optimization of the form (3). Compared to the direct optimization on design D, Gaussianization provides two key advantages. First, the optimization now becomes generic nonlinear programming on the correlation elliptope, which can be efficiently solved using first-order local algorithms. Second, once an optimizer Σ^* is obtained, treatment assignments can be sampled directly via $D_i = g(T_i)$, $T \sim \mathcal{N}(0, \Sigma^*)$. Notably, this Gaussianization idea has been applied in optimization and theoretical computer science literature, where it is known as randomized hyperplane rounding [Williamson and Shmoys, 2011]. More specifically, Goemans and Williamson [1995] propose an approximation algorithm for the Max-Cut problem, where the idea is to generate a cut vector by thresholding a correlated Gaussian vector, with the correlation matrix obtained as the solution to a semidefinite program. Our approach shares a similar procedure when $K = 2$: the action of Gaussianization in our approach is precisely the

Goemans-Williamson rounding, with the analytic function $f(x) = \arccos(x)$ (derived from f_k in Proposition 1) shared in the analysis.

Constraints on Gaussianized design optimization. From the discussion above, the Gaussianization framework focuses on a specific class of designs whose variance-covariance matrices satisfy

$$\mathrm{Cov}_k(D) \in \{f_k(\Sigma) \mid \Sigma \in \mathcal{E}\} .$$

This design class, induced by Gaussianization, is generally a subset of all possible designs, which may be limited under certain scenarios. However, this limitation may still be preferable to the global design optimization that involves NP-hard instances and significant sampling challenges. Probably due to this reason, existing works on design optimality usually focus on specific design classes, such as stratified designs [Bai, 2022] and rerandomization designs [Li and Ding, 2020, Wang and Li, 2023]. We further discuss the connection between Gaussianized designs and existing designs in Section F.

Since the design objective is generally non-convex in Σ, only a local optimizer can be guaranteed. While the local optimality may appear restrictive, it is worth noting that one can flexibly initialize the design optimization from any $\Sigma \in \mathcal{E}$. For example, in the simulation below, we initialize the design optimization using a stratified design. In this sense, our approach serves as a general optimization tool to refine an input design, where the design may already adapt to covariates through other existing methods.

Simulation. To demonstrate the concrete benefits of design optimization, we conduct a simple simulation that evaluates the MSE of $\widehat{\tau}$ under various designs. Two covariate structures are considered: (a) the first covariate has the largest scale and serves as the sole informative feature, and (b) all covariates are uniformly generated and are equally informative. We initialize our iterative algorithm using either an i.i.d. design ($\Sigma = I_n$) or a block design, where Σ is a block correlation matrix representing a classical block design constructed by sorting the first covariate. Further details of the simulation are provided in Section A.

Figure 1 shows the MSE trajectories over iterations in Gaussianized design optimization. In setup (a), initializing with a block design yields a smaller MSE by leveraging the highly informative covariate. Furthermore, starting from the i.i.d. design and minimizing the covariate balance measure with the operator norm results in a lower final MSE. Different initial designs in setup (b) produce similar early-stage MSEs but diverge in their final performance. Notably, with suitable choices of the initial design and the norm, Gaussianized design optimization reduces the MSE by more than 60% through iterations. Figure 2 provides heatmaps of the correlation matrices for different initializations and norms. Observe that Panels (b), (c), (g), and (h) preserve the block structure, which highlights how design optimization makes local improvements.

2 Related Work

Randomized experiments, such as i.i.d. Bernoulli designs and complete randomizations, are generally viewed as robust designs and are thus desirable in practice. Fisher first demonstrated that complete randomization ensures unbiased estimations and therefore facilitates inference and testing [Fisher, 1925, 1926]. Subsequently, Wu [1981] framed robustness in terms of the worst-case mean squared error (MSE) by showing that complete randomization is a minimax design, meaning it minimizes the worst-case MSE. In time-series experiments such as N-of-1 trials, complete randomization has also been shown to be robust against both estimand choices and model misspecifications [Liang and Recht, 2023]. See also Bai [2023], Basse et al. [2023], Harshaw et al. [2019], Kallus [2018], Nordin and Schultzberg [2022] for related discussions.

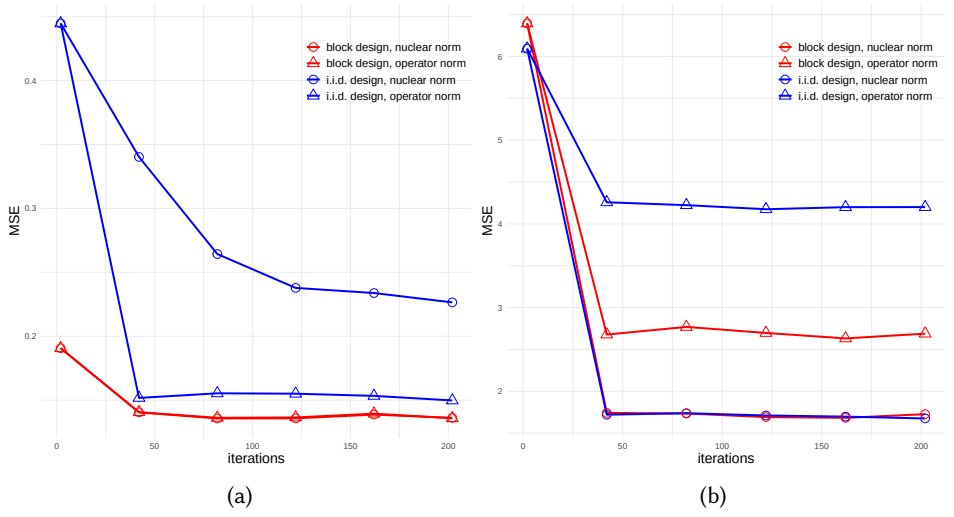


Fig. 1. MSE of the Horvitz-Thompson estimators over iterations of design optimization.

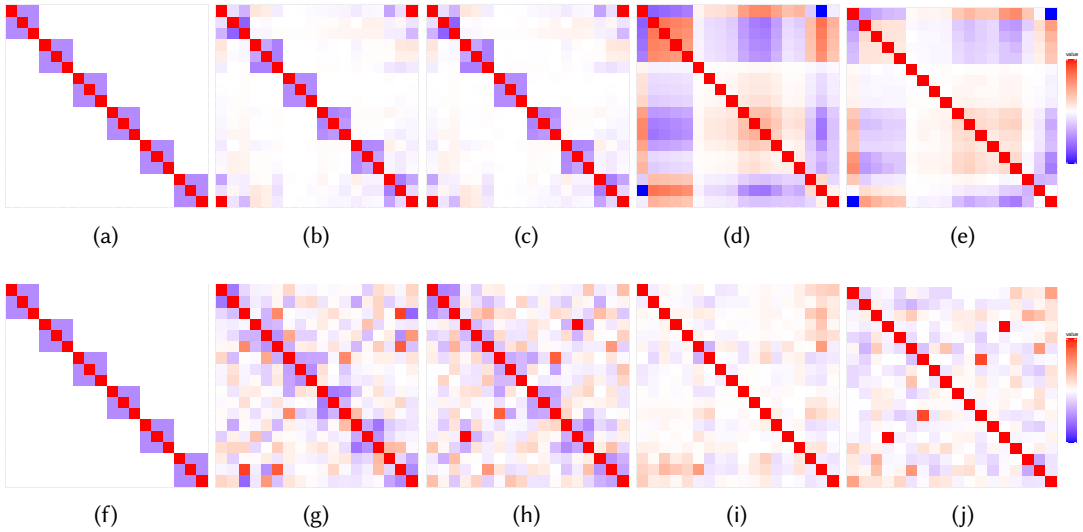


Fig. 2. Heatmaps of covariance matrix Σ from Gaussianized design optimization. The first and second row correspond to the single feature setup and the uniform covariate setup. In each row, from left to right, we show the initial block design, optimized block designs under the operator and nuclear norm, optimized i.i.d. designs under the operator and nuclear norm. Red denotes 1, blue denotes -1, and white denotes 0.

In settings where covariate information is available, which is the focus of this paper, it is reasonable to adapt the experimental design so as to achieve covariate balance across treatment arms. These designs are collectively known as *covariate-adaptive randomization*. With few covariates, blocking is the canonical way to reduce unwanted variation and increase precision [Fisher, 1926]. Matched-pair designs [Greevy et al., 2004, Imbens and Rubin, 2015] are prime examples of blocking, where each block contains two units, and are optimal under certain conditions [Bai, 2022]. However, blocking can be impractical with many covariates. This has motivated sampling-based techniques such as rerandomization [Morgan and Rubin, 2012], which follow an accept-reject

sampling procedure according to certain covariate balance criteria. As shown in Li and Ding [2020], Wang and Li [2023], rerandomization procedures reduce estimation variance by adjusting for the linear component in outcomes that covariates can explain, making them optimal given appropriate covariate balance criteria. However, choosing the right trade-off between covariate balance criteria and the computational complexity of sampling can be challenging, especially with high-dimensional covariates. Moreover, blocking and rerandomization mainly focus on binary treatment settings, and thus it remains unclear how to optimally balance covariates with multiple treatments.

More relevant to our work, Harshaw et al. [2019] introduced the Gram-Schmidt Walk (GSW) design that formalizes the trade-off between covariate balance and robustness in binary treatment settings. Specifically, considering Z as a binary treatment assignment vector, Harshaw et al. [2019] proposed $\|\mathrm{Cov}(Z)\|_{\mathrm{op}}$ and $\|X^\top \mathrm{Cov}(Z)X\|_{\mathrm{op}}$ as measures of robustness and covariate balance, respectively. The GSW design navigates the robustness-balance trade-off by proposing a weighted combination of the aforementioned measures, and employs a random walk to sequentially generate treatment assignments that optimize the weighted combination. Their measure of covariate balance (i.e., $\|X^\top \mathrm{Cov}(Z)X\|_{\mathrm{op}}$) motivates us to study similar norm-based objectives; when $K = 2$, our objective reduces to the GSW objective.

A recent line of work has addressed inference under the covariate-adaptive designs described above, which can be challenging due to the complex covariate-treatment dependencies. See, for instance, Bai et al. [2024], Bugni et al. [2018, 2019], Ma et al. [2020] for inference under stratified designs; Bai et al. [2022] for matched-pair designs; and Li et al. [2018, 2020] for rerandomization. Our asymptotic results under Gaussianization follow this line of work with different proof techniques. In addition to asymptotic inference, we also develop Fisherian-style randomization inference under Gaussianized designs. Although Fisherian randomization was originally developed to test the sharp null of no treatment effects [Fisher, 1935], it has recently been extended to detect heterogeneity [Ding et al., 2016] and interference [Basse et al., 2019, Huang et al., 2025]; these extensions can also be combined with flexible machine learning models for higher efficiency [Guo et al., 2025]. Our paper leverages randomization to compute design-based confidence intervals.

Our paper also contributes to the growing literature on continuous treatments. Using techniques from doubly robust methodology, Colangelo and Lee [2020], Kennedy et al. [2017] studied the estimation of the average potential outcome function, while Hsu et al. [2024] tested functional properties such as monotonicity. Recently, Callaway et al. [2024] analyzed difference-in-differences setups with a continuous treatment, and discussed the identification of response functions and their derivatives. See de Chaisemartin et al. [2022], Dong and Lee [2023], Schindl et al. [2024] for related studies. However, all these works consider i.i.d. data from observational studies, which is distinct from our experimental design setup.

3 A Gaussianization Framework

In this section, we formally introduce the Gaussianization framework, which includes both norm-based covariate balance measures and their Gaussianized representations. Our formulation accommodates general experimental setups with the discrete support $\mathbb{D} = \{1, \ldots, K\}$. We conclude this section by presenting Mehler's formula [Liang and Tran-Bach, 2022, Mehler, 1866], a key technical insight that motivates our design optimization.

3.1 General Covariate Balance Measures

We consider the potential outcome framework as in Section 1.1, and focus on uniform designs such that $\mathbb{P}(D_i = k) = 1/K$ for any $i = 1, \ldots, n$ and $k = 1, \ldots, K$. Non-uniform designs, where D_i follows non-uniform marginal treatment probabilities, can also fit within our Gaussianization framework

by slightly adjusting the Gaussianization function g. The key requirement is that all treatment assignments share the same marginal distribution to enable effective design optimization.

We define our estimand as follows

$$\tau_w = \sum_{k=1}^{K} w_k \tau_k \ , \quad \tau_k = \frac{1}{n} \sum_{i=1}^{n} Y_i(k) \ ,$$

where $w = (w_1, \ldots, w_k)$ is a pre-specified vector. This can be a contrast vector, e.g., $w = (1, -1, 0, \ldots, 0)$, leading to the average treatment effect of treatment arm 1 over 2. It can also be a weight vector, e.g., $w_k = 1/K$ and $\sum w_k = 1$, which reduces to the estimand in Section 1.1 given $K = 3$. These estimands encompass a rich class of causal quantities, and thus they are of primary interest in empirical research.

To estimate τ_w, we use the Horvitz-Thompson estimator as mentioned in Section 1.1:

$$\widehat{\tau}_w = \sum_{k=1}^{K} w_k \widehat{\tau}_k \ , \quad \widehat{\tau}_k = \frac{K}{n} \sum_{i=1}^{n} \mathbb{I}\{D_i = k\} Y_i \ .$$

We focus on Horvitz-Thompson estimators, similar to previous works in the design optimality literature [Bai, 2022, Harshaw et al., 2019, Wang and Li, 2023]. More importantly, the Horvitz-Thompson estimator $\widehat{\tau}_w$ is the optimal linear unbiased sampling estimator of τ_w [Hege, 1967], and thus is desirable for design optimization. Alternatively, one could consider covariate-adjusted estimators [Chang, 2023, Fisher, 1935, List et al., 2024], but their performance is model-specific, potentially leading to biased estimations [Freedman, 2008]. More detailed discussions are provided in Section E.

While $\widehat{\tau}_w$ is unbiased, its mean squared error (MSE) would depend on specific design structures through the covariance matrix of D. The following result has been proved in many works, e.g., Chang [2023].

Lemma 1. *Under uniform experimental designs, for $k = 1, \ldots, K$, we have*

$$\mathbb{E}(\widehat{\tau}_k - \tau_k)^2 = \frac{K^2}{n^2} Y(k)^\top \mathrm{Cov}_k(D) Y(k) \ ,$$

where $Y(k) = (Y_1(k), \ldots, Y_n(k))^\top$, and $\mathrm{Cov}_k(D)$ is defined in Section 1.1.

From Lemma 1, the MSE of k-th treatment effect is a quadratic form in the covariance matrix of the treatment assignment vector, $\mathrm{Cov}_k(D)$, evaluated at the (unknown) potential outcome vector $Y(k)$. Then, for a general estimator $\widehat{\tau}_w$, we utilize the AM-QM inequality to obtain

$$\begin{aligned} \mathbb{E}(\widehat{\tau}_w - \tau_w)^2 &= \mathbb{E}\left(\sum_{k=1}^{K} w_k (\widehat{\tau}_k - \tau_k) \right)^2 \\ &\le K \sum_{k=1}^{K} w_k^2 \mathbb{E}(\widehat{\tau}_k - \tau_k)^2 = \frac{K^3}{n^2} \sum_{k=1}^{3} w_k^2 Y(k)^\top \mathrm{Cov}_k(D) Y(k) \ . \end{aligned}$$

The MSE bound on $\widehat{\tau}_w$ leads to measures of covariate balance. Specifically, following a similar idea as in Harshaw et al. [2019], let's assume for the moment that potential outcomes are perfectly linear in the covariates, i.e., $Y(k) = X\beta_k$, for some $\beta_k \in \mathbb{R}^d$. This reduces the MSE bound to

$$\mathrm{MB} \coloneqq \frac{K^3}{n^2} \sum_{k=1}^{K} w_k^2 \beta_k^\top X^\top \mathrm{Cov}_k(D) X \beta_k \ .$$

In practice, even if we can somehow justify perfect linearity, the signals $\{\beta_k\}_{k=1}^K$ are in general unknown. Harshaw et al. [2019] formulate a worst-case MSE by assuming that the signal has a fixed norm with arbitrary directions. Following their idea, we consider a structural assumption that for $k = 1, \ldots, K$, $\|\beta_k\| \le M$, leading to a measure of worst-case MSE:

$$\sup_{\|\beta_k\| \le M} \mathrm{MB} \propto \sup_{\|\beta_k\| \le M} \sum_{k=1}^K w_k^2 \beta_k^\top X^\top \mathrm{Cov}_k(D) X \beta_k \propto \sum_{k=1}^K w_k^2 \sup_{\|\beta_k\| \le 1} \beta_k^\top X^\top \mathrm{Cov}_k(D) X \beta_k$$
$$= \sum_{k=1}^K w_k^2 \|X^\top \mathrm{Cov}_k(D) X\|_{\mathrm{op}} \,. \tag{4}$$

As an alternative, Isaki and Fuller [1982] and Chang [2023] have introduced the notion of "anticipated variance" that measures an averaged MSE under a prior distribution on the potential outcomes. Following their idea, we consider that $\{\beta_k\}_{k=1}^K$ are random signals with mean zero and identity covariance. This leads to a measure of average-case MSE:

$$\mathbb{E}_{\beta_k} \mathrm{MB} \propto \sum_{k=1}^K w_k^2 \mathbb{E} \beta_k^\top X^\top \mathrm{Cov}_k(D) X \beta_k = \sum_{k=1}^K w_k^2 \,\mathrm{tr}\big(X^\top \mathrm{Cov}_k(D) X \mathbb{E} \beta_k \beta_k^\top\big)$$
$$= \sum_{k=1}^K w_k^2 \,\mathrm{tr}\big(X^\top \mathrm{Cov}_k(D) X\big) = \sum_{k=1}^K w_k^2 \|X^\top \mathrm{Cov}_k(D) X\|_{\mathrm{nuc}} \,. \tag{5}$$

The derivation above leads to the formal definition of covariate balance measures.

Definition 1. *For a uniform design with K treatments, we define the covariate balance measure in the nuclear and operator norm as*

$$\sum_{k=1}^K w_k^2 \|X^\top \mathrm{Cov}_k(D) X\|_{\mathrm{norm}} \,, \quad \mathrm{norm} \in \{\mathrm{nuc}, \mathrm{op}\} \,.$$

To summarize, by making various structural assumptions on the value of β_k, we derive covariate balance measures that only depend on X and the design D. Importantly, this motivates the study of objective (2) as the basis for optimal experimental design, as we show in the sequel. By setting $w_k = 1/K$ and $K = 3$, one recovers the measures exemplified in Procedure 1. Additionally, when $K = 2$, our measure $\sum_{k=1}^K \|X^\top \mathrm{Cov}_k(D) X\|_{\mathrm{op}}$ is equivalent to the covariate balance measure studied in [Harshaw et al., 2019].

3.2 Gaussianized Representation

We introduce a Gaussianized representation of the uniform design D through a map $g : \mathbb{R} \to \{1, \ldots, K\}$ as defined below:

$$g(t) = \begin{cases} i & \text{if } t \in \left(\Phi^{-1}\left(\frac{i-1}{K}\right), \Phi^{-1}\left(\frac{i}{K}\right)\right] \,, \quad i = 1, \ldots, K-1 \\ K & \text{if } t \in \left(\Phi^{-1}\left(\frac{K-1}{K}\right), \infty\right) \end{cases} ,$$

where $\Phi(\cdot)$ is the standard normal CDF. In other words, we discretize the Gaussian treatments T according to the equidistance quantiles. This recovers the uniform design since $g(T_i)$ is uniformly distributed on $\{1, \ldots, K\}$.

When a uniform design is from a Gaussianized representation, i.e., $D_i = g(T_i)$ for $T \sim \mathcal{N}(0, \Sigma)$, the variance-covariance matrix of D is completely captured by Σ. Surprisingly, one can link these two covariance matrices through analytical formulas.

Proposition 1. *Under Gaussianization $D_i = g(T_i)$ for K treatment arms, we have $\mathrm{Cov}_k(D) = f_k(\Sigma)$, where $f_k : [-1,1] \to \mathbb{R}$, $k = 1,\dots,K$ are elementwise functions defined by*

$$f_k(\rho) = \begin{cases} r_{1,1}(\rho) & \text{if } k = 1 \\ r_{K-1,K-1}(\rho) & \text{if } k = K \\ r_{k-1,k-1}(\rho) + r_{k,k}(\rho) - 2r_{k-1,k}(\rho) & \text{otherwise} \end{cases} .$$

For $i, j = 1,\dots,K-1$, we have

$$r_{i,j}(\rho) \coloneqq \mathrm{Cov}(\mathbb{I}\{X \le q_i\}, \mathbb{I}\{Y \le q_j\}) = \int_0^\rho \frac{1}{2\pi\sqrt{1-r^2}} \exp\left(-\frac{q_i^2 + q_j^2 - 2rq_iq_j}{2(1-r^2)}\right) \mathrm{d}r \,, \tag{6}$$

where $q_i = \Phi^{-1}(i/K)$, and (X, Y) follows the bivariate normal distribution with variance one and correlation ρ.

In Proposition 1, the equation $\mathrm{Cov}_k(D) = f_k(\Sigma)$ describes an *elementwise* operation: letting $C = \mathrm{Cov}_k(D)$, it states that $C_{ij} = f_k(\Sigma_{ij})$, $i, j = 1,\dots,n$. This result provides a concrete procedure to compute the covariance matrix $\mathrm{Cov}_k(D)$. Importantly, it facilitates design optimization under Gaussianization, since one can formulate the covariate balance measures as objective functions on Σ:

$$\sum_{k=1}^K w_k^2 \|X^\top \mathrm{Cov}_k(D) X\|_{\mathrm{norm}} = \sum_{k=1}^K w_k^2 \|X^\top f_k(\Sigma) X\|_{\mathrm{norm}} \,. \tag{7}$$

In summary, we propose general covariate balance measures for uniform designs and derive their explicit Gaussianized representations. This Gaussianization enables feasible design optimization algorithms over the space of Gaussian covariance matrices, which will be the focus of Section 4.

Proposition 1 warrants more technical clarifications. First, its main benefit comes from (6), which provides analytical expressions of $\mathrm{Cov}_k(D)$. Alternatively, one may evaluate each covariance in (6) by Monte Carlo, but such simulation-based methods can be computationally challenging for large-scale randomized experiments. Second, we illustrate below the shape of f_k through the three-treatment example.

Remark 1 (Evaluation of f_k in the three-treatment example). *Given $K = 3$, we evaluate $f(\rho) = \sum_{k=1}^3 f_k(\rho)$, which maps Σ to $\sum_{k=1}^3 \mathrm{Cov}_k(D)$. This function represents the design optimization objective in Section 1.1, since for $w_k = 1/3$, the covariate balance measure in the nuclear norm reduces to*

$$\sum_{k=1}^3 w_k^2 \|X^\top \mathrm{Cov}_k(D) X\|_{\mathrm{nuc}} \propto \Big\|X^\top \sum_{k=1}^3 \mathrm{Cov}_k(D) X\Big\|_{\mathrm{nuc}} = \big\|X^\top f(\Sigma) X\big\|_{\mathrm{nuc}} \,.$$

From the visualization of $f(\cdot)$ in Figure 3, we observe that negative (positive resp.) correlations in Σ induce negative (positive resp.) correlations in $\sum_k \mathrm{Cov}_k(D)$, with $f(0) = 0$ being a fixed point. More interestingly, $f(-1)$ and $f(0)$ induce similar correlations that are close to zero, implying that perfect negative correlation and zero correlation in T lead to similar MSE performance. Lastly, we highlight that the derivative of f goes to infinity at the endpoints ± 1. This singular behavior of f' will guide us in developing optimization algorithms in Section 4.

3.3 Mehler's Formula and Proof Sketch of Proposition 1

We prove Proposition 1 by leveraging a representation [Liang and Tran-Bach, 2022] of bivariate normal distribution based on Mehler's formula and Hermite polynomials. To begin with, we define Hermite polynomials in the probabilists' convention.

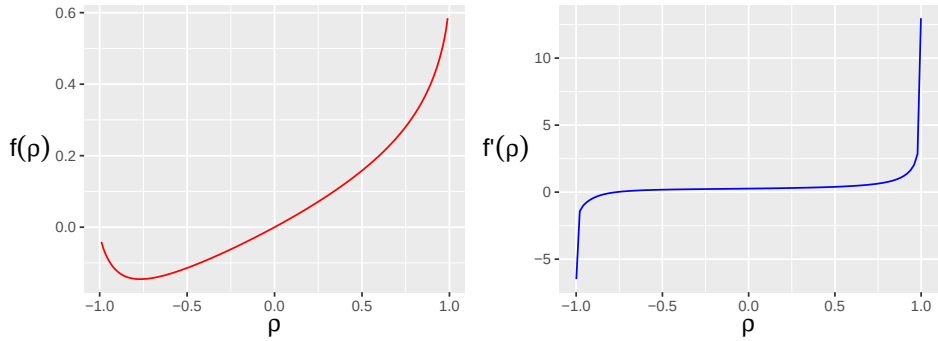


Fig. 3. Function $f(\rho)$ and its derivative $f'(\rho)$ on $(-1, 1)$.

Definition 2. *For non-negative integers $m \geq 0$, the m-th order Hermite polynomial is defined by*

$$\mathrm{He}_m(x) = \frac{(-1)^n}{\phi(x)} \frac{\mathrm{d}^m}{\mathrm{d}x^m} \phi(x) \ .$$

Here ϕ is the standard normal density function. Define the normalized Hermite polynomials as

$$h_m(x) := \frac{1}{\sqrt{m!}} \mathrm{He}_m(x) \ .$$

Let L^2_ϕ be the class of square-integrable functions with respect to the standard normal distribution. Then, the set $\{h_m\}_{m=0}^{\infty}$ forms an orthonormal basis of L^2_ϕ as one can verify that

$$\mathbb{E}_{Z\sim\mathcal{N}(0,1)} \left[h_m(Z) h_{m'}(Z)\right] = \mathbb{I}\{m = m'\} \ .$$

We can then define the Hermite coefficients as follows.

Definition 3. *For any function $g \in L^2_\phi$, the m-th Hermite coefficient is defined by*

$$\alpha_m[g] := \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(0,1)} \left[g(Z) h_m(Z)\right] \ .$$

Let $p_\rho(x, y)$ be the density function of the bivariate normal distribution with variance one and correlation ρ. Mehler's formula [Mehler, 1866] connects $p_\rho(x, y)$ to Hermite polynomials as shown below:

$$p_\rho(x, y) = \sum_{m=0}^{\infty} \rho^m h_m(x) h_m(y) \phi(x) \phi(y) \ .$$

That is, the density $p_\rho(x, y)$ can be decomposed into a sequence of products of Hermite polynomials and standard normal densities. Based on this result, we establish a representation for the covariance of functions defined over bivariate normal distributions.

Lemma 2. *For $g, h \in L^2_\phi$, if $(X, Y) \in \mathbb{R}^2$ follow a bivariate normal distribution with variance one and correlation ρ, we have*

$$\mathop{\mathrm{Cov}}_{(X,Y)} \left[g(X), h(Y)\right] = \sum_{m=1}^{\infty} \alpha_m[g] \alpha_m[h] \rho^m \ .$$

Based on Mehler's formula and Lemma 2, we show a sketch proof for Proposition 1. A complete proof can be found in Section B.1.

Proof Sketch. Here we focus on the proof of Equation (6), which is the key step in proving the result. Let $g(x) = \mathbb{I}\{x \le q_i\}$ and $h(x) = \mathbb{I}\{x \le q_j\}$. We leverage the derivative representation of Hermite polynomials (Definition 2) to obtain

$$\alpha_m[g] = -\frac{1}{\sqrt{m!}}\phi(q_i)\,\mathrm{He}_{m-1}(q_i)\,, \quad \alpha_m[h] = -\frac{1}{\sqrt{m!}}\phi(q_j)\,\mathrm{He}_{m-1}(q_j)\,.$$

Based on Lemma 2, this implies

$$r_{ij}(\rho) = \sum_{m=1}^{\infty} \frac{1}{m!}\,\mathrm{He}_{m-1}(q_i)\,\mathrm{He}_{m-1}(q_i)\phi(q_i)\phi(q_j)\rho^m\,.$$

Notice that $r_{ij}(0) = 0$ and

$$r'_{ij}(\rho) = \sum_{m=1}^{\infty} \frac{1}{(m-1)!}\,\mathrm{He}_{m-1}(q_i)\,\mathrm{He}_{m-1}(q_i)\phi(q_i)\phi(q_j)\rho^{m-1} = p_\rho(q_i, q_j)\,.$$

We obtain

$$r_{ij}(\rho) = \int_0^\rho p_r(q_i, q_j)\mathrm{d}r = \int_0^\rho \frac{1}{2\pi\sqrt{1-r^2}}\exp\left(-\frac{q_i^2 + q_j^2 - 2rq_iq_j}{2(1-r^2)}\right)\mathrm{d}r\,.$$

□

In summary, Proposition 1 can be proved by applying Mehler's formula to the covariance expression in (6). This trick will be used again in design optimization for the continuous setting (Section 5). Notably, this technical tool is designed for bivariate normal distributions, which further motivates the act of Gaussianization of treatments.

4 Gaussianized Design Optimization

In this section, we will focus on solving the following optimization problems:

$$\min_{\Sigma\in\mathcal{E}} \|X^\top f(\Sigma)X\|_{\mathrm{norm}} =: l_{\mathrm{norm}}(\Sigma)\,, \mathrm{norm} \in \{\mathrm{nuc}, \mathrm{op}\}\,, \tag{8}$$

where f is a given elementwise function defined on $[-1, 1]$. Based on the linearity of the nuclear norm, the objective in (8) in $\|\cdot\|_{\mathrm{nuc}}$ is equivalent to (7) by setting $f(\rho) = \sum_k w_k^2 f_k(\rho)$. Under the operator norm, the design optimization problem (7) is a weighted sum of objectives in the form of (8), and one can slightly modify the algorithm below to solve (7). Moreover, the general optimization problem (8) encompasses other covariate balance objectives in Section 5.

Formally, we propose Algorithm 1 to solve (8) above. This algorithm applies projected gradient descent (PGD-Gauss) on a factorized representation of Σ, similar to the Burer-Monteiro approach in semidefinite programming [Burer and Monteiro, 2003].

The function f in design optimization may have an infinite derivative at ±1 (Remark 1). Conceptually, this type of f will set ±1 to be a barrier. Therefore, in Algorithm 1, we fix the diagonal values of Σ^t and only update on the off-diagonal entries. That is, we consider

$$\nabla l_{\mathrm{nuc}}(\Sigma^t) = (XX^\top - \mathrm{diag}(XX^\top)) \circ f'(V^tV^{t\top})\,,$$
$$\nabla l_{\mathrm{op}}(\Sigma^t) = (Xu_1u_1^\top X^\top - \mathrm{diag}(Xu_1u_1^\top X^\top)) \circ f'(V^tV^{t\top})\,,$$

where $\circ$ is the Hadamard product, $u_1 \in \mathbb{R}^d$ is the leading eigenvector of $X^\top f(\Sigma^t)X$. For diagonal elements in the gradient, we adopt the convention $0 \times f'(\pm1) = 0$. Notably, f' can be obtained by directly differentiating the analytic functions f_k defined in Proposition 1. That is, Proposition 1 enables the direct computation of the gradient in Gaussianized design optimization.

ALGORITHM 1: Projected Gradient Descent for Gaussianized Design Optimization (PGD-Gauss)

Data: $X \in \mathbb{R}^{n\times d}$, an evaluation function f, and an initial design Σ^1. Number of iterations T.
Result: Optimized covariance matrix Σ^*.
begin
  Parametrize $\Sigma^1 = V^1(V^1)^\top$, where $V^1 \in \mathbb{R}^{n\times k}$, $\|v_i^1\| = 1$, and k equals to the rank of Σ^1. Here v_i^1 is the i-th row of V^1.
  for $t = 1, \ldots, T$ **do**
    Compute $\nabla l_{\text{norm}}(\Sigma^t)$.
    $V^{t+1} = \left[I_n - \eta_t \nabla l_{\text{norm}}(\Sigma^t)\right] V^t$ for a proper step size η_t.
    $v_i^{t+1} \leftarrow v_i^{t+1}/\|v_i^{t+1}\|$.
    $\Sigma^{t+1} \leftarrow V^{t+1}(V^{t+1})^\top$.
  end
  $\Sigma^* \leftarrow \Sigma^T$.
end

Since the objective function is non-convex in Σ in general, the PGD-Gauss only obtains a local optimizer near the initial covariance matrix Σ^1. As explained in Section 1.1, Gaussianized design optimization is not tailored to identify the global solution that perfectly balances the covariates, but rather to serve as a tool for achieving local improvements based on a given input design. In Section D, we further discuss a numerical approach to quantify the performance gap between the PDG-Gauss solution and the global optimizer.

By default, we initialize the design optimization by setting $\Sigma^1 = I_n$, which results in i.i.d. treatments. We view this as the baseline Gaussian design, as it does not take any covariate information, and i.i.d. designs have a robust performance against unknown outcome-generating models [Harshaw et al., 2019]. Therefore, the number of steps we run PGD-Gauss is an explicit tradeoff between robustness and covariate balance. In simulations of Section 7, the i.i.d. initialization leads to satisfactory performance compared to state-of-the-art designs.

As the optimization problem (8) can be cast as a general optimization on the correlation manifold, one may consider implementing other algorithms designed for optimization problems over manifolds [Boumal, 2014]. In our paper, we focus on the PGD-Gauss algorithm, since it is a first-order algorithm and is more computationally efficient compared to other second order methods. We discuss more details about its computational complexity in Section A.5.

5 Gaussian Design with Continuous Treatments

In this section, we extend Gaussianization to settings with continuous treatments. Specifically, we introduce a new experimental design, called Gaussian design, to give continuous treatments based on multivariate Gaussian distribution.

Definition 4 (Gaussian Design). *A Gaussian design allocates treatment T_i to unit i, where $T = (T_1, \ldots, T_n) \sim \mathcal{N}(0, \Sigma)$ for some $\Sigma \in \mathcal{E}$.*

When the actual treatment is restricted to a bounded interval $[a, b]$, one may compute a rescaled treatment assignment $(a + b)/2 + T_i(b - a)/(2z_{0.999})$, where z_α denotes the α-quantile of the standard normal distribution. This ensures that the rescaled treatment falls in $[a, b]$ with high probability (> 0.998). Gaussian designs directly allocate continuous treatments as above, and it is thus mechanically different from the Gaussianization perspective, where we focus on discrete treatments but model them using latent Gaussian variables. Compared to Gaussianization, Gaussian designs capture average structural properties of potential outcome functions as we discuss below.

5.1 Causal Estimands

We denote $Y_i(t)$ to be the response function for unit i and $t \in \mathbb{R}$, which generalizes the potential outcomes to continuous treatments. With an abuse of notation, we use $Y_i = Y_i(T_i)$ to denote the observed outcome for unit i. Given continuous treatments and response functions, we work with a class of causal effects of the form

$$\tau_w^c = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} Y_i(t)w(t)\phi(t)\mathrm{d}t\ , \tag{9}$$

where $w(\cdot)$ is a pre-specified weight function on different treatment values. We use the superscript c in τ_w^c to distinguish it from τ_w under the discrete setting.

Similar to the discrete setup, we focus on Horvitz-Thompson-type estimators

$$\widehat{\tau}_w^c = \frac{1}{n}\sum_{i=1}^{n} Y_i(T_i)w(T_i) = \frac{1}{n}\sum Y_i W_i\ , \quad W_i \coloneqq w(T_i)\ .$$

Clearly, $\widehat{\tau}_w^c$ is an unbiased estimator of τ_w^c under Gaussian design. In the following, we provide several leading examples of the weight function $w(\cdot)$ in (9) to get meaningful causal estimands.

Example 1 (Average Treatment Effects on a Given Interval). *Suppose we want to learn about the average treatment effect on a treatment interval $[r, l]$ [Fryges and Wagner, 2008]. We may set $w(t) = \frac{\mathbb{I}\{t\in[r,l]\}}{(l-r)\phi(t)}$, which leads to*

$$\tau_w^c = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{l-r}\int_r^l Y_i(x)\mathrm{d}x\ , \quad \widehat{\tau}_w^c = \frac{1}{n}\sum_{i=1}^{n} Y_i\frac{\mathbb{I}\{T_i\in[r,l]\}}{(l-r)\phi(T_i)}\ .$$

Example 2 (First derivative). *Suppose $Y_i(t)$ is differentiable with $\mathbb{E}Y_i^2(T_i) < 0$ and $\mathbb{E}|Y_i'(T_i)| < 0$. To learn the first derivative of response functions, we consider $w(t) = t$ and obtain*

$$\tau_w^c = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} Y_i(t)t\phi(t)\mathrm{d}t \overset{(i)}{=} \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} Y_i'(t)\phi(t)\mathrm{d}t\ , \quad \widehat{\tau}_w^c = \frac{1}{n}\sum_{i=1}^{n} Y_i T_i\ ,$$

where (i) follows from Stein's Lemma. Notably, if we replace the base Gaussian density $\phi(t)$ with $\psi(t) = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$, the causal estimand reduces to $\tau_w^c = \frac{1}{2n}\sum_{i=1}^{n}(Y_i(1) - Y_i(-1))$, which resembles the average treatment effect in binary treatment setups.

Example 3 (Second derivative). *Suppose $Y_i(t)$ is twice differentiable with $\mathbb{E}Y_i^2(T_i) < 0$ and $\mathbb{E}|Y_i''(T_i)| < 0$. To learn the second derivative, we consider $w(t) = t^2 - 1$ and obtain*

$$\tau_w^c = \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} Y_i(t)(t^2-1)\phi(t)\mathrm{d}t \overset{(i)}{=} \frac{1}{n}\sum_{i=1}^{n}\int_{\mathbb{R}} Y_i''(t)\phi(t)\mathrm{d}t\ , \quad \widehat{\tau}_w^c = \frac{1}{n}\sum_{i=1}^{n} Y_i(T_i^2-1)\ ,$$

where (i) follows from an extension of Stein's Lemma [Mamis, 2022]. $\widehat{\tau}_w^c$ serves as an unbiased estimator for the average second derivative of response functions.

5.2 Variance Formula and Measures of Covariate Balance

To get traction on estimating the variance of the estimators, we decompose $Y_i(t)$ as follows:

$$Y_i(t) = a_i Y_0(t) + b_i\ , \quad a_i = X_i^\top\beta_1\ , \quad b_i = X_i^\top\beta_2\ . \tag{10}$$

In this decomposition, a_i and b_i control the scale and location of the i-th response function, and they are perfectly linear in covariates. $Y_0(t)$ is a baseline response function, which is assumed to be known by the researcher. This assumption is justified as researchers often have prior knowledge of

the shape of response functions, such as sigmoid dose-response curves in clinical trials [Meddings et al., 1989], and exponential utility functions in economics [Arrow, 1971].

Under (10), we analyze the variance of $\widehat{\tau}_w^c$. For two random vectors $X, Y \in \mathbb{R}^d$, we use the notation $\mathrm{Cov}(X, Y) := \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)^\top]$ and $\mathrm{Cov}(X) := \mathrm{Cov}(X, X)$. Then, one can show that under Equation (10), it holds that

$$\begin{aligned}\mathrm{Var}(\widehat{\tau}_w^c) &= \frac{1}{n^2}\left(\beta_1^\top X^\top \mathrm{Cov}(Y_0 \circ W) X\beta_1 + \beta_2^\top X^\top \mathrm{Cov}(W) X\beta_2 + 2\beta_1^\top X^\top \mathrm{Cov}(Y_0 \circ W, W) X\beta_2\right) \\ &\le \frac{2}{n^2}\left(\beta_1^\top X^\top \mathrm{Cov}(Y_0 \circ W) X\beta_1 + \beta_2^\top X^\top \mathrm{Cov}(W) X\beta_2\right) . \end{aligned} \tag{11}$$

With a slight abuse of notation, we define $Y_0 = (Y_0(T_1), \ldots, Y_0(T_n))^\top$, $W = (W_1, \ldots, W_n)^\top$, $\circ$ is the Hadamard (elementwise) product, and the second line follows from the AM-GM inequality. From Equation (11), the estimation performance is characterized by quadratic forms similar to the discrete setting (Lemma 1). In addition, the variance in inequality (11) depends on the coefficients β_1, β_2, which are unknown in general.

To make progress, we adopt a similar approach as in Section 3. We first assume that β_1, β_2 are random signals with mean zero and identity covariance matrix, which leads to a measure of average-case MSE:

$$\begin{aligned}\mathop{\mathbb{E}}_{\beta_1,\beta_2} \mathrm{Var}(\widehat{\tau}_w^c) &\le \frac{2}{n^2}\,\mathrm{tr}\big(X^\top(\mathrm{Cov}(Y_0 \circ W) + \mathrm{Cov}(W))X\big) \\ &\propto \|X^\top \mathrm{Cov}(Y_0 \circ W) X\|_{\mathrm{nuc}} + \|X^\top \mathrm{Cov}(W) X\|_{\mathrm{nuc}} .\end{aligned}$$

Alternatively, by assuming $\|\beta_1\| \le M$, $|\beta_2\| \le M$, we obtain an upper bound on the worst-case MSE:

$$\begin{aligned}\sup_{\|\beta_1\|\le M, \|\beta_2\|\le M} \mathrm{Var}(\widehat{\tau}_w^c) &\le \sup_{\|\beta_1\|\le M, \|\beta_2\|\le M} \frac{2}{n^2}\left(\beta_1^\top X^\top \mathrm{Cov}(Y_0 \circ W) X\beta_1 + \beta_2^\top X^\top \mathrm{Cov}(W) X\beta_2\right) \\ &\propto \|X^\top \mathrm{Cov}(Y_0 \circ W) X\|_{\mathrm{op}} + \|X^\top \mathrm{Cov}(W))X\|_{\mathrm{op}} .\end{aligned}$$

These analytical steps lead to the formal definition of covariate balance measures under Gaussian design in the continuous setting.

Definition 5. *For Gaussian designs with a baseline response function Y_0 and a weight function w, define the average and worst-case covariate balance measures as*

$$\|X^\top \mathrm{Cov}(Y_0 \circ W) X\|_{\mathrm{norm}} + \|X^\top \mathrm{Cov}(W) X\|_{\mathrm{norm}} , \mathrm{norm} \in \{\mathrm{nuc}, \mathrm{op}\} . \tag{12}$$

5.3 Gaussianized Representation

Using Mehler's formula and Hermite coefficients in Section 3, we derive the following result, which is a direct application of Lemma 2.

Proposition 2. *Suppose that $Y_0 w : t \mapsto Y_0(t)w(t) \in L^2_\phi$ and $w \in L^2_\phi$. Then we have*

$$\mathrm{Cov}(Y_0 \circ W) = f_{Y_0, w}(\Sigma) , \quad \mathrm{Cov}(W) = f_w(\Sigma) .$$

Here, $f_{Y_0,w}$ and f_w are elementwise functions defined by

$$f_{Y_0,w}(\rho) = \sum_{m=1}^{\infty} \alpha_m[Y_0 w]^2 \rho^m , \quad f_w(\rho) = \sum_{m=1}^{\infty} \alpha_m[w]^2 \rho^m , \quad \rho \in [-1, 1] ,$$

where $\alpha_m[g]$ is the m-th Hermite coefficient of the function g.

Proposition 2 demonstrates that the covariance matrices in covariate balance measures can be explicitly written as functions of Σ. This result facilitates optimization over Σ, similar to the role of Proposition 1 in the uniform design setup. Combining the results above, we formulate covariate balance measures

$$\|X^\top f_{Y_0,w}(\Sigma)X\|_{\text{norm}} + \|X^\top f_w(\Sigma)X\|_{\text{norm}}\,, \text{norm} \in \{\text{nuc}, \text{op}\}\ .$$

Consequently, one may directly apply the algorithm proposed in Section 4 to Gaussian design and optimize the covariate balance. We evaluate this design concretely in Section A.4.

6 Asymptotics and Inference

In this section, we study asymptotic properties and inference under the Gaussianization $T \sim \mathcal{N}(0, \Sigma)$, where Σ is a solution obtained from the PGD-Gauss algorithm in Section 4, within the design-based framework. In design-based inference [Imbens and Rubin, 2015], we view the potential outcomes as fixed and the only randomness comes from the treatment assignment, i.e., the Gaussian treatment T. Here, we prove asymptotic normality under Gaussianization, and provide concrete procedures to compute confidence intervals. The key takeaway is that Gaussianization under the PGD-Gauss solution results in smaller variance compared to i.i.d. Gaussianization, and thus improves estimation efficiency. Notably, our asymptotic theory allows high-dimensional covariates, where d can grow with, or even be larger than n.

For presentation purposes, we focus on the uniform design setup in Section 3 where the treatments are modeled by $D_i = g(T_i)$. Similar results for continuous treatments are discussed in Section B.2.3.

6.1 Asymptotic Normality

Here, we focus on the PGD-Gauss under the nuclear norm objective $\|X^\top f_k(\Sigma)X\|_{\text{nuc}}$. Recall that f_k defined in Proposition 1 is the covariance mapping with respect to treatment k, and thus this objective serves as a covariate balance measure for k-th average treatment effect. Similar normality results can be shown under the operator norm, but we will focus on the nuclear norm for simplicity.

We study the asymptotic properties of the average treatment effect for arm k:

$$\widehat{\tau}_k = \frac{K}{n}\sum_{i=1}^{n} Y_i \mathbb{I}\{D_i = k\}\,, \quad D_i = g(T_i)\ .$$

We focus on implementing one step of PGD-Gauss with step size η using the default initialization $\Sigma^1 = I_n$, and denote the obtained solution by Σ_η. We impose the following assumption on the step size η and covariates X.

Assumption 1. *The covariates $X \in \mathbb{R}^{n\times d}$ satisfy $\|X_i\| = 1$, i.e., each row of X has unit norm. The step size in PGD-Gauss satisfies $\eta\|XX^\top - I_n\|_{\text{op}} = o(1)$.*

Assumption 1 requires that Σ_η is a local perturbation of I_n by controlling the step size, which is the key condition to establish asymptotic normality. To better understand the step size condition, we may consider $X_i \overset{iid}{\sim} \mathcal{N}(0, \frac{1}{d}I_d)$, so that $\|X_i\| \approx 1$ in expectation. Then, the random matrix theory suggests that $\|XX^\top\|_{\text{op}} = O(n/d)$ with high probability [Tropp et al., 2015]. If, for intuition, we assume that $n > d$, the step size condition boils down to $\eta = o\left(\frac{n}{d}\right)$.

To characterize the asymptotic distribution under the one-step PGD-Gauss, we define a sequence of ancillary potential outcomes. Specifically, using the f_k in Proposition 1, we define

$$\tilde{Y}(k) = f_k(I_n)^{-1/2} f_k(\Sigma_\eta)^{1/2} Y(k)\ .$$

Theorem 1. *Suppose Assumption 1 holds. Consider Gaussianization $T \sim \mathcal{N}(0, \Sigma_\eta)$, where Σ_η is the obtained solution from the one-step PGD-Gauss. If, as n goes to infinity,*

(1) $\max_{i=1,\dots,n} \tilde{Y}_i^2(k)/n \to 0$,
(2) $n\,\mathrm{Var}(\widehat{\tau}_k) = \frac{K-1}{n}\sum_{i=1}^n \tilde{Y}_i^2(k)$ *has a positive limit,*
(3) $\|Y(k)\|^2 \le nM$ *for some constant* $M > 0$,

it holds that

$$\sqrt{n}\,(\widehat{\tau}_k - \tau_k) \xrightarrow{d} \mathcal{N}(0, \lim_{n\to\infty} n\,\mathrm{Var}(\widehat{\tau}_k)) = \mathcal{N}\left(0, \lim_{n\to\infty} \frac{K-1}{n}\|\tilde{Y}(k)\|^2\right) .$$

Theorem 1 establishes the asymptotic normality of $\widehat{\tau}_k$ under Gaussianization. The proof of Theorem 1 relies on the asymptotic equivalence between $\widehat{\tau}_k$ under Σ_η and an ancillary estimator under i.i.d. Gaussianization. Once the asymptotic equivalence is established, it suffices to prove the asymptotic normality for the ancillary estimator using Lindeberg's central limit theorem. Notably, Condition (1) is a Lindeberg-type condition, while Condition (2) ensures a non-degenerate limiting distribution. Both assumptions are common in the asymptotic theory of design-based inference [Li and Ding, 2017]. A full proof and a generalization to multi-step PGD-Gauss can be found in Section B.2.

Notably, the variance term in Theorem 1 indicates the benefit of running PGD-Gauss for covariate balance. To see this, we may define

$$V(\Sigma) \coloneqq \frac{K-1}{n}\|\tilde{Y}(k)\|^2 .$$

From the proof of Theorem 1, $V(\Sigma)$ not only captures the variance limit, but also exactly matches the finite-sample variance of $\sqrt{n}(\widehat{\tau}_k - \tau_k)$, i.e., the MSE of $\widehat{\tau}_k$ after rescaling. The following proposition shows that $V(\Sigma_\eta)$ is strictly smaller than $V(I_n)$ on average. Denote by $\|A\|_F$ the Frobenius norm of a matrix A. Write $a_n = \Omega(b_n)$ if there exists a constant c such that $a_n \ge cb_n$ for n large enough.

Proposition 3. *Suppose* $Y(k) = X\beta_k^\top$, *where* β_k *is a random signal with zero mean and identity covariance matrix. In addition, suppose Assumption 1 holds and*

$$\eta = o(1) , \quad n\eta^3\|XX^\top - I_n\|_{\mathrm{op}}^4 = o(\|XX^\top - I_n\|_F^2) .$$

If $f_k'(0) \neq 0$, *we have*

$$\mathbb{E}_{\beta_k} V(I_n) - \mathbb{E}_{\beta_k} V(\Sigma_\eta) = \Omega\left(\frac{\eta}{n}\|XX^\top - I_n\|_F^2\right) > 0 .$$

Proposition 3 suggests that for n large enough, there is a nonzero improvement in $V(\Sigma_\eta)$ in the average sense above. Therefore, $\widehat{\tau}_k$ has a smaller variance under Σ_η compared to the initial design I_n, which reveals the benefit of covariate balance. However, we clarify that the improvement in Proposition 3 is with respect to the non-asymptotic variance $V(\Sigma_\eta)$, which does not directly translate into an improvement in the limiting variance of the asymptotic distribution. Theoretical conditions under which the one-step PGD-Gauss reduces the limiting variance remain an open and complex problem, which we consider as future work.

6.2 Inference

As with asymptotic normality, the validity of inferential procedures depends on the structure of the covariance matrix, and hence it is difficult to devise a single method that applies to all Gaussianizations. Here we distinguish between two cases: for $\Sigma = I_n$ (i.i.d. Gaussianization), we propose explicit variance estimators with normality-based confidence intervals; for general Σ with more complex covariance structures, we introduce a randomization-based confidence interval. As in previous section, we focus on the uniform design setup (Section 3) and the estimator $\widehat{\tau}_k$.

Under i.i.d. Gaussianization, Theorem 1 suggests that

$$\sqrt{n}\,(\widehat{\tau}_k - \tau_k) \xrightarrow{d} \mathcal{N}\left(0, \lim_{n\to\infty} \frac{K-1}{n}\|Y(k)\|^2\right) .$$

To conduct inference, one needs to estimate the asymptotic variance $V_{iid} \coloneqq \frac{K-1}{n}\sum_i Y_i^2(k)$. We propose a Horvitz-Thompson-type variance estimator

$$\widehat{V}_{iid} = \frac{K-1}{n}\sum_{i=1}^{n} K Y_i^2 \mathbb{I}\{g(T_i) = k\} . \tag{13}$$

Theorem 2. *Suppose that* $\max_i |Y_i(k)| = O(1)$. *We have* $\mathbb{E}\widehat{V}_{iid} = V_{iid}$ *and* $\mathrm{Var}(\widehat{V}_{iid}) = O(1/n)$. *Hence* $\widehat{V}_{iid}$ *is a consistent estimator of* V_{iid}.

Combining Theorems 1 and 2, we derive the design-based confidence interval

$$[\widehat{\tau}_k - z_{\alpha/2}\sqrt{\widehat{V}_{iid}/n}\ ,\ \widehat{\tau}_k + z_{\alpha/2}\sqrt{\widehat{V}_{iid}/n}] .$$

Here, we set $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$, resulting in an asymptotic $(1-\alpha)$ confidence interval for $\widehat{\tau}_k$. For a general estimator $\widehat{\tau}_w$ that involves multiple treatment arms, consistent variance estimation is more challenging, as the asymptotic variance V_{iid} would depend on covariances between different potential outcomes, which are generally unobservable. In those cases, one can use a conservative variance estimator, which we discuss in Section B.2.3.

For general Gaussian covariance matrices, we propose a randomization-based confidence interval as below. This can be viewed as a variant of parametric bootstrap. Recall that T_i and Y_i denote the observed treatment and outcome for unit i, respectively.

Procedure 2 (Randomization-Based Confidence Interval for $\widehat{\tau}_k$).

(1) *For* $k = 1, \ldots, K-1$, *fit a model* $\widehat{m}_k(X_i)$ *by regressing* Y_i *over* X_i *for all units with treatment* k.
(2) *Generate* $\{T^b\}_{b=1}^B \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$. *For each randomization* T^b, *impute outcomes for units receiving treatment* k *by*

$$Y_i^b = \begin{cases} Y_i & \text{if } T_i^b = T_i \\ \widehat{m}_k(X_i) & \text{if } T_i^b \neq T_i \end{cases} .$$

Compute the randomization-based estimate $\widehat{\tau}_k^b$ *based on* T_i^b *and* Y_i^b, $i = 1, \ldots, n$.
(3) *Construct the randomization-based confidence interval* $[\widehat{c}(\alpha/2)\ ,\ \widehat{c}(1-\alpha/2)]$, *where* $\widehat{c}(\alpha)$ *is the* α*-sample quantile for* $\{\widehat{\tau}_k^b\}_{b=1}^B$.

Procedure 2 conducts simulation-based inference by first learning a regression model to impute all potential outcomes under treatment k, and then generating new treatments and outcomes to simulate the distribution of the estimator $\widehat{\tau}_k$. The validity of Procedure 2 hinges on step 1, i.e., how well the fitted model captures the true outcome functions, which will be numerically validated in Section A. Additionally, one can apply Procedure 2 to compute the confidence interval for general estimators $\widehat{\tau}_w$ in Section 3 by replacing $\widehat{\tau}_k^b$ with $\widehat{\tau}_w^b$ in step 2.

Conceptually, Procedure 2 follows similar ideas as Imbens and Menzel [2018], which introduce a causal bootstrap to construct confidence intervals for the average treatment effect. However, Imbens and Menzel [2018] focus exclusively on a binary treatment setting under complete randomization, whereas Procedure 2 accommodates multiple treatment arms and general Gaussianized designs.

7 Simulations

Here we conduct comprehensive experiments on different designs under a factorial setup. Additional numerical results can be found in Section A, including a real data example in the continuous treatment setting and simulation details on the 3-treatment setup in Section 1.1.

We set $n = 100$, $d = 5$, and $X_i \overset{iid}{\sim} \mathcal{N}(0, I_d)$. Consider a factorial design with two treatments $A_i \in \{0, 1\}$, $B_i \in \{0, 1\}$ with potential outcomes: $Y_i(A_i, B_i) = X_i^\top \beta_1 + A_i(X_i^\top \beta_2) + B_i(0.2 + X_i^\top \beta_3) + 0.5 A_i B_i + \varepsilon_i$, where $\beta_1 = (-1, -1, -2/3, -6/5, 0)$, $\beta_2 = (0, 0, -8/5, 8/5, 8/5)$, $\beta_3 = (2, 2, 2, 0, 0)^\top$, $\varepsilon_i \sim \mathcal{N}(0, 0.1^2)$, and we fix ε_i for different potential outcomes. To translate the factorial design to a standard uniform design, we encode the treatments by $D_i = 1 + 2A_i + B_i \in \{1, 2, 3, 4\}$. Then, we apply the Gaussianization techniques in Section 3 to model the treatments by $D_i = g(T_i)$ for the map g in Section 3, enabling Gaussianized design optimization.

In the factorial design under the potential outcome framework, one is usually interested in estimating main effects and interaction effects [Dasgupta et al., 2015]:

$$\tau_1 \coloneqq \frac{1}{2n} \sum_{i=1}^{n} (-Y_i(0,0) - Y_i(0,1) + Y_i(1,0) + Y_i(1,1)) = 0.25 + \frac{1}{n} \sum_{i=1}^{n} X_i^\top \beta_2 \,,$$

$$\tau_2 \coloneqq \frac{1}{2n} \sum_{i=1}^{n} (-Y_i(0,0) + Y_i(0,1) - Y_i(1,0) + Y_i(1,1)) = 0.45 + \frac{1}{n} \sum_{i=1}^{n} X_i^\top \beta_3 \,,$$

$$\tau_{12} \coloneqq \frac{1}{2n} \sum_{i=1}^{n} (Y_i(0,0) - Y_i(0,1) - Y_i(1,0) + Y_i(1,1)) = 0.25 \,.$$

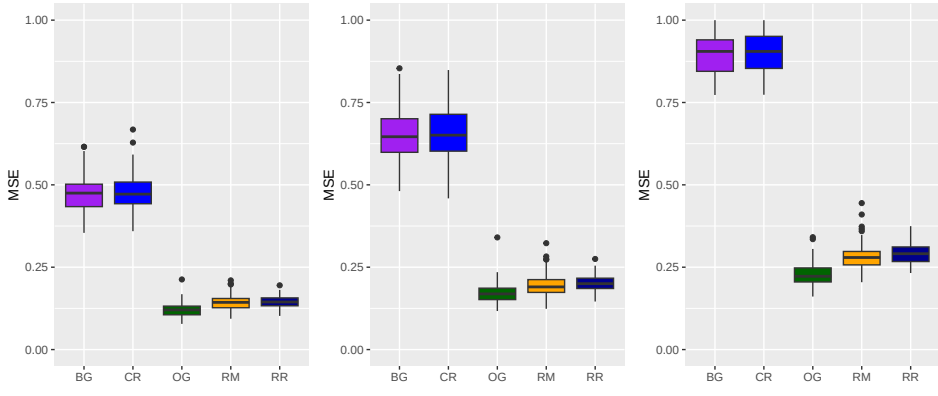
We will estimate these quantities based on Horvitz-Thompson estimators.

We evaluate the MSE of Horvitz-Thompson estimators under different designs. We implement baseline Gaussianization (BG) with $\Sigma = I_n$, and the optimized Gaussianization (OG) with Σ^*. The optimized covariance matrix Σ^* is obtained from PGD-Gauss for solving the nuclear-norm objective with i.i.d. initialization and 200 iterations. For comparison purposes, we implement complete randomization (CR) [Dasgupta et al., 2015], recursive matching (RM) [Bai et al., 2024] and rerandomization (RR) [Li et al., 2020]. RM and RR can be considered as state-of-the-art designs for covariate balance in the factorial setup.

The MSEs are presented in boxplots in Figure 4, where we evaluate the MSEs based on 1,000 simulations and generate different covariates and potential outcomes for 100 times. We observe across all three estimation problems, OG achieves the smallest MSE among five designs. In Figure 5, we provide a scatter plot of the MSEs for τ_1 over the covariate balance objective in the nuclear norm, evaluated under all different designs. We observe that 1) a smaller covariate balance measure indicates smaller MSE on average, and 2) OG achieves the smallest covariate balance measure across all designs, trailed by RM and RR. In Section A, we provide further simulation details including design-based confidence intervals and robustness checks under different outcome models.

8 Conclusion

In our paper, we develop a Gaussianization framework to optimize experimental designs for covariate balance. This approach accommodates general covariates and multiple treatment arms, offering a key advantage over existing methods. Moreover, Gaussianization seamlessly extends to continuous treatments via the Gaussian design, which may be of independent interest in practical applications. As an extension, it would be interesting to consider more complex settings, such as those involving interference. Second, developing a general asymptotic theory for Gaussianized designs that extends beyond local perturbations remains an open problem. We consider these areas promising topics for future work.



(a) τ_1 (b) τ_2 (c) τ_{12}

Fig. 4. MSEs for estimating $\tau_1, \tau_2, \tau_{12}$ under different designs.

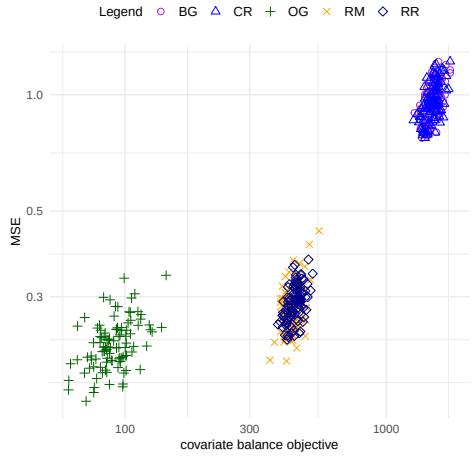


Fig. 5. Scatter plot of MSEs over the covariate balance objective (nuclear norm) evaluated under different designs.

Acknowledgments

This research is supported by the NSF Career Award DMS-2042473 and by the Wallman Society of Fellows at the University of Chicago.

References

Kenneth J Arrow. 1971. The theory of risk aversion. *Essays in the theory of risk-bearing* 90 (1971), 120.

Yuehao Bai. 2022. Optimality of matched-pair designs in randomized controlled trials. *American Economic Review* 112, 12 (2022), 3911–3940.

Yuehao Bai. 2023. Why randomize? Minimax optimality under permutation invariance. *Journal of Econometrics* 232, 2 (2023), 565–575.

Yuehao Bai, Jizhou Liu, and Max Tabord-Meehan. 2024. Inference for matched tuples and fully blocked factorial designs. *Quantitative Economics* 15, 2 (2024), 279–330.

Yuehao Bai, Joseph P Romano, and Azeem M Shaikh. 2022. Inference in experiments with matched pairs. *J. Amer. Statist. Assoc.* 117, 540 (2022), 1726–1737.

Francisco Barahona and Ali Ridha Mahjoub. 1986. On the cut polytope. *Mathematical programming* 36 (1986), 157–173.

Guillaume Basse, Avi Feller, and Panos Toulis. 2019. Randomization tests of causal effects under interference. *Biometrika* 106, 2 (02 2019), 487–494. doi:10.1093/biomet/asy072 arXiv:https://academic.oup.com/biomet/article-pdf/106/2/487/28575447/asy072.pdf

Guillaume W Basse, Yi Ding, and Panos Toulis. 2023. Minimax designs for causal effects in temporal experiments with treatment habituation. *Biometrika* 110, 1 (2023), 155–168.

Allan Borodin and Ran El-Yaniv. 2005. *Online computation and competitive analysis*. cambridge university press.

Nicolas Boumal. 2014. Optimization and estimation on manifolds. (2014).

Federico A Bugni, Ivan A Canay, and Azeem M Shaikh. 2018. Inference under covariate-adaptive randomization. *J. Amer. Statist. Assoc.* 113, 524 (2018), 1784–1796.

Federico A Bugni, Ivan A Canay, and Azeem M Shaikh. 2019. Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics* 10, 4 (2019), 1747–1785.

Samuel Burer and Renato DC Monteiro. 2003. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical programming* 95, 2 (2003), 329–357.

Brantly Callaway, Andrew Goodman-Bacon, and Pedro HC Sant'Anna. 2024. *Difference-in-differences with a continuous treatment*. Technical Report. National Bureau of Economic Research.

Haoge Chang. 2023. Design-based Estimation Theory for Complex Experiments. arXiv:2311.06891 [econ.EM]

Kyle Colangelo and Ying-Ying Lee. 2020. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036* (2020).

Tirthankar Dasgupta, Natesh S Pillai, and Donald B Rubin. 2015. Causal inference from 2K factorial designs by using potential outcomes. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 77, 4 (2015), 727–753.

Clément de Chaisemartin, Xavier d'Haultfoeuille, Félix Pasquier, and Gonzalo Vazquez-Bare. 2022. Difference-in-differences estimators for treatments continuously distributed at every period. *arXiv preprint arXiv:2201.06898* (2022).

Peng Ding, Avi Feller, and Luke Miratrix. 2016. Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 78, 3 (2016), 655–671.

Yingying Dong and Ying-Ying Lee. 2023. Nonparametric Doubly Robust Identification of Causal Effects of a Continuous Treatment using Discrete Instruments. *arXiv preprint arXiv:2310.18504* (2023).

Pascaline Dupas. 2014. Short-Run Subsidies and Long-Run Adoption of New Health Products: Evidence From a Field Experiment. *Econometrica* 82, 1 (January 2014), 197–228.

R.A. Fisher. 1925. *Statistical methods for research workers*. Edinburgh Oliver & Boyd.

Ronald A Fisher. 1926. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* 33 (1926), 503–513.

Roland A. Fisher. 1935. *The Design of Experiments*. Oliver and Boyd, Edinburgh.

David Freedman. 2008. On Regression Adjustments to Experimental Data. *Advances in Applied Mathematics* 40 (02 2008), 180–193. doi:10.1016/j.aam.2006.12.003

Helmut Fryges and Joachim Wagner. 2008. Exports and productivity growth: First evidence from a continuous treatment approach. *Review of World Economics* 144 (2008), 695–722.

Alan S Gerber and Donald P Green. 2012. Field experiments: Design, analysis, and interpretation. *(No Title)* (2012).

Michel X Goemans and David P Williamson. 1995. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)* 42, 6 (1995), 1115–1145.

Robert Greevy, Bo Lu, Jeffrey H Silber, and Paul Rosenbaum. 2004. Optimal multivariate matching before randomization. *Biostatistics* 5, 2 (2004), 263–275.

Wenxuan Guo, JungHo Lee, and Panos Toulis. 2025. ML-assisted Randomization Tests for Detecting Treatment Effects in A/B Experiments. arXiv:2501.07722 [stat.ME] https://arxiv.org/abs/2501.07722

Christopher Harshaw, Fredrik Sävje, Daniel Spielman, and Peng Zhang. 2019. Balancing covariates in randomized experiments with the Gram-Schmidt Walk design.

Vijaya S Hege. 1967. An Optimum Property of the Horvitz-Thomson Estimate. *J. Amer. Statist. Assoc.* 62, 319 (1967), 1013–1017.

Yu-Chin Hsu, Martin Huber, Ying-Ying Lee, and Chu-An Liu. 2024. Testing monotonicity of mean potential outcomes in a continuous treatment with high-dimensional data. *Review of Economics and Statistics* (2024), 1–44.

Shunzhuang Huang, Xinran Li, and Panos Toulis. 2025. Randomization Tests for Monotone Spillover Effects. arXiv:2501.02454 [stat.ME] https://arxiv.org/abs/2501.02454

Mark Huber and Nevena Maric. 2017. Bernoulli correlations and cut polytopes. *arXiv preprint arXiv:1706.06182* (2017).

Guido Imbens and Konrad Menzel. 2018. *A causal bootstrap*. Technical Report. National Bureau of Economic Research.

Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Cary T. Isaki and Wayne A. Fuller. 1982. Survey Design Under the Regression Superpopulation Model. *J. Amer. Statist. Assoc.* 77, 377 (1982), 89–96. http://www.jstor.org/stable/2287773

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

Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. 2017. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 79, 4 (2017), 1229–1245.

EL Lehmann and HJ D'Abrera. 1975. *Nonparametrics: Statistical methods based on ranks.* Holden-Day.

Xinran Li and Peng Ding. 2017. General forms of finite population central limit theorems with applications to causal inference. *J. Amer. Statist. Assoc.* 112, 520 (2017), 1759–1769.

Xinran Li and Peng Ding. 2020. Rerandomization and regression adjustment. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82, 1 (2020), 241–268.

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

Xinran Li, Peng Ding, and Donald B Rubin. 2020. Rerandomization in 2 K factorial experiments. *The Annals of Statistics* 48, 1 (2020), 43–63.

Tengyuan Liang and Benjamin Recht. 2023. Randomization inference when n equals one. *arXiv preprint arXiv:2310.16989, Biometrika, forthcoming* (2023).

Tengyuan Liang and Hai Tran-Bach. 2022. Mehler's formula, branching process, and compositional kernels of deep neural networks. *J. Amer. Statist. Assoc.* 117, 539 (2022), 1324–1337.

Winston Lin. 2013. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics* 7, 1 (2013), 295 – 318.

John A List, Ian Muir, and Gregory Sun. 2024. Using machine learning for efficient flexible regression adjustment in economic experiments. *Econometric Reviews* 44, 1 (2024), 2–40.

Wei Ma, Yichen Qin, Yang Li, and Feifang Hu. 2020. Statistical inference for covariate-adaptive randomization procedures. *J. Amer. Statist. Assoc.* 115, 531 (2020), 1488–1497.

Konstantinos Mamis. 2022. Extension of Stein's lemma derived by using an integration by differentiation technique. *Examples and Counterexamples* 2 (2022), 100077.

JB Meddings, RB Scott, and GH Fick. 1989. Analysis and comparison of sigmoidal curves: application to dose-response data. *American Journal of Physiology-Gastrointestinal and Liver Physiology* 257, 6 (1989), G982–G989.

F.G. Mehler. 1866. Ueber die Entwicklung einer Function von beliebig vielen Variablen nach Laplaceschen Functionen höherer Ordnung. *Journal für die reine und angewandte Mathematik* (1866), 161–176.

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

Jerzy Neyman. 1923. On the application of probability theory to agricultural experiments: essay on principles (with discussion); section 9 (in Polish). (1923). Engl. transl. by D. M. Dabrowska and T. P. Speed (1990), Statist. Sci., 5, 465–472.

Mattias Nordin and Mårten Schultzberg. 2022. Properties of restricted randomization with implications for experimental design. *Journal of Causal Inference* 10, 1 (2022), 227–245.

Kyle Schindl, Shuying Shen, and Edward H. Kennedy. 2024. Incremental effects for continuous exposures. arXiv:2409.11967 [stat.ME] https://arxiv.org/abs/2409.11967

Joel A Tropp et al. 2015. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning* 8, 1-2 (2015), 1–230.

Yuhao Wang and Xinran Li. 2023. Asymptotic Theory of the Best-Choice Rerandomization using the Mahalanobis Distance. arXiv:2312.02513 [stat.ME] https://arxiv.org/abs/2312.02513

David P Williamson and David B Shmoys. 2011. *The design of approximation algorithms.* Cambridge university press.

Chien-Fu Wu. 1981. On the robustness and efficiency of some randomized designs. *The Annals of Statistics* (1981), 1168–1177.

A Additional Simulation Results

A.1 Simulation Details of the 3-treatment Experiment

For the example in Section 1.1, we consider the following setup.

- $n = 18, d = 5$.
- For covariates, we consider (a) single feature: $X_{i1} \sim \mathcal{N}(2, 3^2)$ and $X_{ij} \sim \mathcal{N}(0, 0.1^2)$ for $j = 2, \ldots, d$. $\beta_{1k} \sim 2 + 2\exp(1)$ and $\beta_{jk} \sim 2\exp(1)$; (b) uniform covariates: $X_{ij} \sim \mathcal{N}(0, 3.6^2)$ and $\beta_{jk} \sim 2\exp(1)$. Here, β_{jk} denotes the j-th entry of the vector β_k for $j \in \{1, \ldots, d\}$ and $k \in \{1, 2, 3\}$, and $\exp(1)$ is the exponential distribution with the rate parameter equal to one.
- Generate potential outcomes based on $Y(k) = X\beta_k$.

In the block initialization, we first construct the size-3 blocks by sorting the first coordinate of X. Then, for each block matrix, we set diagonals to be 1 and off-diagonal entries to be -0.5. We run the PGD-Gauss in Section 4 for 200 iterations.

A.2 Simulation Details under the Factorial Setup

In CR, one assigns same number of units to different treatments uniformly at random, which serves as a baseline that does not leverage covariate information. In RR, we repeatedly generate treatment assignments from CR according to the covariate balance criteria on Mahalanobis distance with the asymptotic acceptance probability $p_a = 0.01$ as defined in Li et al. [2020, Section 4]. In RM, we recursively match the experimental units for different treatment factors following [Bai et al., 2024].

Here, we follow the simulation setup in Section 7 and compare the computable confidence intervals under different designs. The confidence intervals for BG and OG can be constructed based on Section 6. For OG, we employ Procedure 2 based on fitting a linear model $\widehat{m}_k$ of outcomes over covariates for each treatment arm. For RR and CR, we adopt the variance estimators proposed in [Dasgupta et al., 2015, Li et al., 2020], respectively, and construct confidence intervals based on asymptotics. Here, we exclude the recursive matching design (RM) because, although it is a powerful design, the inferential results in Bai et al. [2022] are derived in a superpopulation framework which is distinct from our design-based framework.

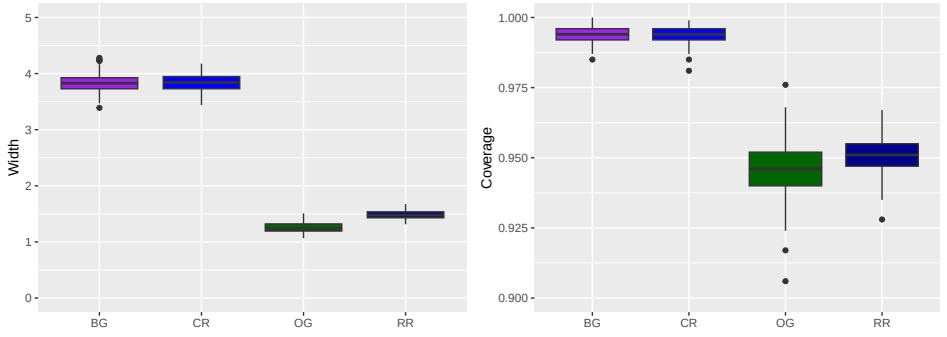
We present in Figure 6 the boxplots of the width of confidence intervals, along with the coverage rates. For simplicity, we focus on τ_1, as the results for τ_2 and τ_{12} are similar. Note that all methods achieve a correct coverage of 95%, while some of them are conservative. In term of the width, we observe that OG returns shortest confidence intervals, which reveals practical benefits of our design.

A.3 Robustness Checks under the Factorial Setup

Here, we check the performance of different designs under more complex outcome generating processes. We follow the factorial setup in Section 7, but consider two covariate structure: (1) i.i.d. covariates $X_{ij} \sim \mathcal{N}(0, 2)$, and (2) clustered covariates $X_{ij} \sim \mathcal{N}(1, 0.5)$ for $i = 1, \ldots, \lfloor n/2 \rfloor$ and $X_{ij} \sim \mathcal{N}(-1, 0.5)$ for $i = \lfloor n/2 \rfloor + 1, \ldots, n$. In addition, we change the outcome model as follows:

$$
\begin{aligned}
Y_i(0,0) &= X_i^\top \beta + \varepsilon_i(0,0), \\
Y_i(0,1) &= X_i^\top \beta + X_{i1}^3 \beta_1 + \varepsilon_i(0,1), \\
Y_i(1,0) &= X_i^\top \beta + X_{i2}^3 \beta_2 + X_{i3}^3 \beta_3 + \varepsilon_i(1,0), \\
Y_i(1,1) &= X_i^\top \beta + \varepsilon_i(1,1),
\end{aligned}
$$

where $\beta \sim \mathcal{N}(0, I_d)$, $\beta_1, \beta_2, \beta_3 \sim \mathcal{N}(0, 1)$, and $\varepsilon_i(A_i, B_i) \sim \mathcal{N}(0, 1^2)$ for $A_i \in \{0, 1\}, B_i \in \{0, 1\}$. That is, we inspect a more complex outcome model with nonlinear components.



(a) Width (b) Coverage

Fig. 6. Width of confidence intervals and their coverage rates for τ_1 under different designs.

Table 1 shows the performance of different designs under the different settings explained above. Each column of the table below shows the (relative) average MSE for a given design compared to the baseline Gaussianization (BG). The BG column is always one by definition, and a smaller ratio indicates better MSE performance. We observe that OG still shows a relatively small MSE across different cases, and maintains the smallest MSE when $d \geq 20$. Note that due to the increasing dimension, the rerandomization design [Li et al., 2020] becomes computationally challenging, and thus we switch the proposed rerandomization design to the best-choice rerandomization [Wang and Li, 2023], which mitigates the computational burden by fixing the number of total rerandomizations. Specifically, we set the number of rerandomization to be 500 for all the simulations in Table 1.

dimension d	BG	CR	OG	RM	RR
Panel A: i.i.d. covariates					
5	1.00	0.98	0.38	0.56	0.98
10	1.00	0.98	0.40	0.68	1.00
20	1.00	1.01	0.48	0.79	0.99
30	1.00	1.00	0.51	0.86	1.00
40	1.00	0.99	0.46	0.90	1.00
50	1.00	1.01	0.51	0.89	1.02
Panel B: clustered covariates					
5	1.00	0.99	0.75	0.37	1.00
10	1.00	1.03	0.64	0.51	1.02
20	1.00	1.01	0.54	0.66	0.99
30	1.00	1.01	0.54	0.78	1.02
40	1.00	1.01	0.46	0.84	1.00
50	1.00	0.99	0.44	0.63	1.00

Table 1. Comparison of relative MSEs across different covariate structures and dimensions.

A.4 Bed Nets Study on Continuous Treatments

In Dupas [2014], the authors conducted a field experiment in Kenya, where households in different regions were encouraged to purchase insecticide-treated bed nets designed to prevent malaria.

Price (in Kenyan Shillings)	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6
0					96.9(64)	98.1 (53)
40		75.4(61)				
50			72.4 (58)	40.0 (35)		
60					73.0(37)	
70	55.2(29)					
80		57.1(70)				
90			55.0(60)			
100	34.0(47)			28.6(49)		61.1(18)
110					32.4(37)	
120		28.1(64)				
130	24.5(49)					
140					37.9(29)	
150			31.0(58)	35.6(45)		22.2 (18)
190	17.9(28)					
200		17.0(59)		10.3(29)		
210			18.8 (48)			
250	6.7(30)			7.7 (26)		

Table 2. Rates at which anti-malaria bed nets are purchased, by sales price (after subtracting the value of a randomly assigned voucher). The total number of households per group is in parentheses, and the exchange rate at the time of this study was 65 shillings = $1.00.

Dupas [2014] treated households by sending vouchers with different discounted prices for the bed nets, effectively inducing a continuous price variable. The original outcome was a binary variable indicating whether a household purchased the bed nets using the voucher, and Dupas [2014], Gerber and Green [2012] analyzed the effect of voucher on purchase rates of bed nets. Here, we implement Gaussian designs to assign continuous price treatments, and evaluate their performance compared with the original design in their study.

The original experiment in Dupas [2014] was a 2-stage randomization (referred to as 2S), which fixes the price variable at discrete levels:

- Stage 1 (region-level): Assign treatment levels (discounted prices) for six different regions in Kenya. These values are fixed once assigned throughout the experiment.
- Stage 2 (household-level): Randomly assign treatments for households in each region, with the treatment levels determined in Stage 1.

The design and results of the bed nets study are presented in Table 2, which reports the proportion of households which purchased a bet net given a region and a discounted price. For instance, in region 2 with price 40, there were 61 households who received the voucher and 75.4% of them eventually redeemed the voucher and purchased bed nets. Clearly, the rate at which bed nets were purchased declines steadily as the price increases: 75.4% of households offered a price of 40 shillings purchased a net, compared to only 17.0% of those offered a price of 200 shillings.

A.4.1 Estimation of Linear Effect. In our numerical study, we define each experimental unit as a cluster of households corresponding to a data point in Table 2, with outcome defined as the proportion of households who purchased bed nets. This results in an experiment on 26 cluster-level units. Each unit i has a dummy covariate vector $U_i \in \mathbb{R}^6$ indicating the region of the unit, and a cluster-level covariate vector $X_i \in \mathbb{R}^3$, including the proportion of male heads sampled to receive the voucher, the proportion of households that have ever shopped at the shop, and the average age of the female heads in households. The three covariates are selected due to their statistical significance in an OLS regression of outcome over all collected covariates.

To assess the performance of different designs, we need to impute the outcome value at any counterfactual price level. To this end, we use the following imputation model:

$$Y_i(t) = X_i^\top \alpha_1 + U_i^\top \alpha_2 + U_i^\top \beta t + \varepsilon_i \,,$$

The coefficients $\alpha_1, \alpha_2, \beta$ are OLS estimates for this linear model based on the observed data, and $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ where σ^2 is the OLS estimate of the error variance. Our goal is to estimate the average linear treatment effect under the imputation model

$$\tau_L^c \coloneqq \frac{1}{n} \sum_{i=1}^{n} U_i^\top \beta \,.$$

Under the original design, one can unbiasedly estimate τ_L^c by

$$\widehat{\tau}_{2S} = \frac{1}{n} \sum_{i=1}^{n} Y_i^{2S} \sum_{j=1}^{6} U_{ij} \frac{D_i^{2S} - \mu_j}{\sigma_j^2}$$

D_i^{2S} in $\widehat{\tau}_{2S}$ denotes the treatment in the 2-stage (2S) design, i.e., randomly selected from the discrete set of price levels for each region. Accordingly, μ_j and σ_j^2 are the mean and variance of D_i^{2S} in region j.

We discuss implementation details about Gaussian design toward estimating τ_L^c. First, since the price treatment takes values in $[0, 250]$, we implement $T_i \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = 125$ and $\sigma = 41.67 = 250/6$, ensuring that that T_i falls in $[0, 250]$ with high probability. Then, noticing that the estimand τ_L^c is same as the average of $Y_i'(t)$, we follow Example 2 to obtain an unbiased estimator

$$\widehat{\tau}_L^c = \frac{1}{n} \sum_{i=1}^{n} Y_i w_L(T_i) \,, \quad w_L(t) = \frac{t - \mu}{\sigma^2} \,.$$

Lastly, to perform Gaussianized design optimization in Section 5, we specify a linear baseline response function

$$Y_0(t) = -\frac{t}{250} + 1 \,. \tag{14}$$

The specified response function captures the true imputation model in the sense that

$$Y_i(t) = a_i Y_0(t) + b_i \,, \quad a_i = -250 U_i^\top \beta \text{ and } b_i = X_i^\top \alpha_1 + U_i^\top \alpha_2 + \varepsilon_i - 250 U_i^\top \beta \,,$$

which validates the modeling assumption in Equation (10). We initialize PGD-Gauss from i.i.d. Gaussian design with covariates $\{(X_i, U_i)\}_{i=1}^{26}$, and obtain the optimized Gaussian design after 200 iterations. We focus on the baseline response function (14), i.i.d. initialization, and the nuclear norm objective in design optimization throughout the bed nets study.

Table 3 presents the MSE and inference properties for different designs. We implement the baseline i.i.d. Gaussian design (BG) and the original 2S design for comparison. To conduct inference, we use the conservative variance estimator for BG and randomization-based confidence intervals

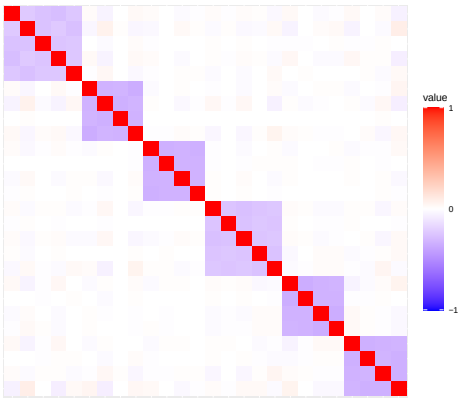


Fig. 7. Heatmap of the optimized Gaussian covariance in OG.

for OG and 2S (Section B.2.3).[1] We observe that OG achieves the smallest MSE as well as the shortest confidence interval. We observe a numerical gap between the actual coverage rates and the expected 95% coverage for BG, which is due to the small sample size ($n = 26$).

design	$n\times$ MSE	average CI width $\times\sqrt{n}$	coverage (%)
$\tau_L^c = -3.75 \times 10^{-3}$			
BG	1.2×10^{-4}	4.00×10^{-2}	90.7
OG	0.5×10^{-4}	2.59×10^{-2}	97.9
2S	0.8×10^{-4}	3.26×10^{-2}	100.0

Table 3. MSE properties and inference for linear effects based on 1,000 simualtions.

We visualize the optimized Gaussian covariance matrix in Figure 7. The covariance matrix — initialized from the identity matrix— automatically learns the block structure for units from regions 1-6 under PGD-Gauss. In addition, within each block, it reveals an approximate equicorrelation structure, which resembles the covariance matrix of complete randomization. In short, the OG design in this setup performs a continuized block randomization.

A.4.2 Testing Monotonicity and Convexity.

Monotonicity. Testing the monotonicity, say, non-decreasingness, can be formulated as the following hypothesis on underlying response functions:

$$H_0^M : Y_i'(t) \geq 0 \,, \text{ for any } i = 1, \ldots, n \text{ and } t \in \mathbb{R}.$$

[1]We apply Procedure 3 by fitting a linear model $\widehat{m}(x, t)$ of outcomes over treatments and all covariates, i.e., $Y_i \sim X_i + U_i + T_i$. Note that the original inference procedure in Dupas [2014] is no longer applicable under our setup, as we are considering a different causal estimand.

Directly testing for H_0^M is impossible, since we have only one observation for each response function. We consider a weaker null hypothesis of H_0^M

$$H_{0,g}^M : \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(\mu,\sigma^2)} Y_i'(Z) \geq 0 \ .$$

This weak null hypothesis is motivated by Gaussian design, and it indicates that the derivative averaged over units and treatments is non-negative. The design-induced hypothesis $H_{0,g}^M$ allows us to check monotonicity through Gaussian design.

Similar to Section A.4.1, we consider an imputation model with a nonlinear component

$$Y_i(t) = X_i^\top \alpha_1 + U_i^\top \alpha_2 + bU_i^\top \beta t^3 + \varepsilon_i \ ,$$

The coefficients $\alpha_1, \alpha_2, \beta$ are OLS estimates for this linear model based on the observed data and $b = 1$, and $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ where σ^2 is the OLS estimate of the error variance. We report that each element of β is negative, indicating that the null hypothesis H_0^M is false. To evaluate the power under different degree of monotonicity, we inspect $b = 0, 0.5, 1, 1.5, 2$, where a larger b indicates more significant decreasingness in the data.

Same as Section A.4.1, under Gaussian design, one can use $\widehat{\tau}_L^c$ to unbiasedly estimate

$$\tau_M^c \coloneqq \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(\mu,\sigma^2)} Y_i'(Z) \ ,$$

which is guaranteed by Example 2. Hence, we implement BG and OG to test for $H_{0,g}^M$ by checking whether the computed confidence interval for τ_M^c is below zero. Confidence intervals are computed in the same way as in Section A.4.1. For comparison purposes, we employ a parametric approach that first fits an OLS regression on

$$Y_i \sim X_i + U_i + T_i \ ,$$

and then applies a t-test for T_i as a surrogate method to check monotonicity. We evaluate the parametric linear model approach (LM) under all three designs BG, OG, and 2S.

From Figure 8(a), OG is more powerful for testing $H_{0,g}^M$ compared to BG, justifying the benefits of covariate balance. Under the LM approach, the original 2S design provides the highest power. However, we note that LM approaches are not directly comparable with BG and OG approach, as they target the null hypothesis that whether the OLS coefficient is negative, which is different from $H_{0,g}^M$.

Convexity. Similar to the monotonicity case, we test $H_0^C : Y_i''(t) > 0$ for any i through a weaker null hypothesis

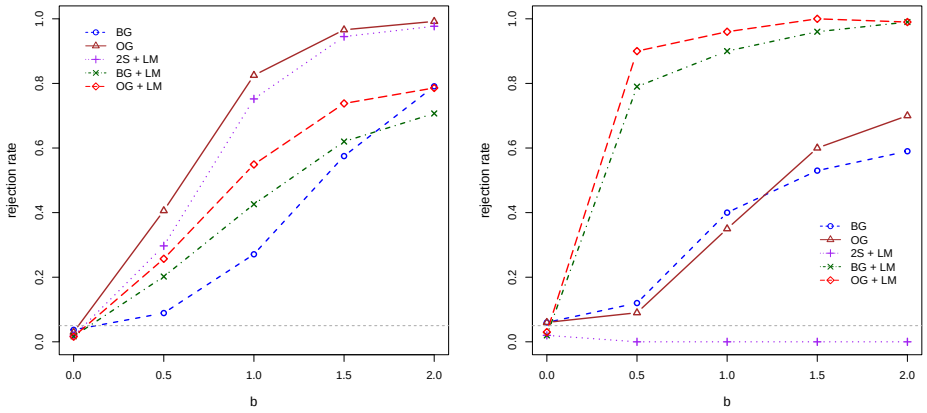
$$H_{0,g}^C : \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{Z\sim\mathcal{N}(\mu,\sigma^2)} Y_i''(Z) \geq 0 \ .$$

We consider an imputation model

$$Y_i(t) = X_i^\top \alpha_1 + U_i^\top \alpha_2 + bU_i^\top \beta t^2 + \varepsilon_i \ .$$

The coefficients $\alpha_1, \alpha_2, \beta$ are OLS estimates for this linear model based on the observed data and $b = 1$, and $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ where σ^2 is the OLS estimate of the error variance. Since each element of β is negative, the imputation model implies that the null hypothesis H_0^C is false, i.e., the response functions are concave. Again, we inspect $b = 0, 0.5, 1, 1.5, 2$, where a larger b indicates more significant concavity in the data.

Convexity reflects the second-order information of the response functions, which typically requires larger sample sizes to gain any meaningful conclusions. Hence, to make nontrivial power



(a) Monotonicity (b) Convexity

Fig. 8. Rejection rates for testing monotonicity and convexity over different b. Rejection means that the confidence interval for the parameter of interest is strictly below zero.

comparisons, we simulate a new set of covariates of size $n = 500$, by sampling uniformly from the original covariates of 26 samples. The covariates are fixed once generated. This ends up with a new experimental setup with 500 units.

Under Gaussian designs, we compute the following estimator

$$\widehat{\tau}_C^c = \frac{1}{n}\sum_{i=1}^{n} Y_i w_C(T_i)\,, \quad w_C(t) = \frac{((t-\mu)^2/\sigma^2 - 1)}{\sigma^2}\,,$$

which is an unbiased estimator of $\tau_C^c \coloneqq \frac{1}{n}\sum \mathbb{E}Y_i''(Z)$ based on Example 3.[2] Hence, we implement BG and OG to test for $H_{0,g}^C$ by checking whether the computed confidence interval for τ_C^c is below zero.[3] For comparison purposes, we implement a parametric approach that fits a linear regression model

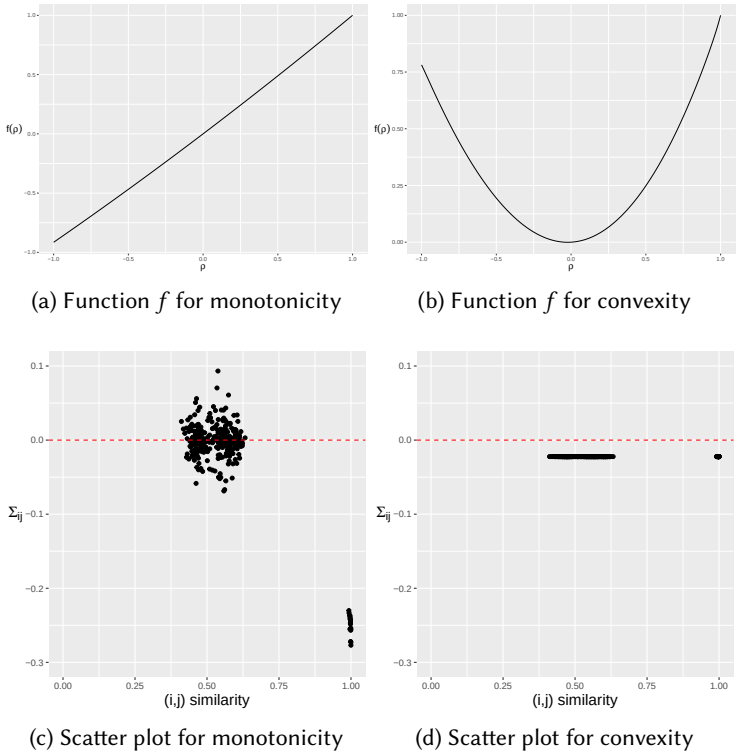
$$Y_i \sim X_i + U_i + T_i + T_i^2$$

and applies the t-test for the coefficient of T_i^2 to check for convexity. We evaluate the parametric linear model approach (LM) under BG, OG, and 2S.

From Figure 8(b), OG and BG have similar performance, and OG achieves higher power only for $b = 1.5, 2$. This is because the optimized covariance matrix for OG is numerically similar to that for BG, the identity matrix, as we will explain below. Among all methods, OG combined with LM (OG + LM) yields highest power. Note that the LM approach under the original design fails to reject convexity, as the 2S design focuses on discrete treatment values, making it difficult to probe the concave structure.

Estimands and Optimized Gaussian Designs. We conclude our numerical study by showing how different estimands lead to different structures in the optimized covariance matrix of OG. In Figure 9(a)-(b), we visualize the function f in the covariate balance objective $\|X^\top f(\Sigma)X\|_{\text{nuc}}$ for

[2]The subscript C denotes the convexity, whereas the superscript c denotes the continuous setting.
[3]In Procedure 3, we fit a linear model $\widehat{m}(x,t)$ based on $Y_i \sim X_i + U_i + U_i T_i$.



(a) Function f for monotonicity (b) Function f for convexity

(c) Scatter plot for monotonicity (d) Scatter plot for convexity

Fig. 9. Function f and the correlation structure across different designs. The red dashed line indicates zero correlation, which corresponds to the i.i.d. Gaussian design.

monotonicity and convexity. That is, based on Section 5, we compute

$$f(\rho) \coloneqq f_{Y_0,w}(\rho) + f_w(\rho)\,, \quad \rho \in [-1, 1]\,,$$

where w corresponds to w_L and w_C defined before, and Y_0 is the linear baseline response function (14). Observe that they are approximately linear and quadratic functions. In the second row, we visualize the scatter plot for the off-diagonal entries Σ_{ij} in the optimized covariance and a pairwise covariate-similarity $X_i^\top X_j / \|X_i\| \|X_j\|$ for all $i \neq j$. In (c), the optimized design balances the covariates by assigning negative correlations to pairs of units with higher similarities. In (d), the optimized covariance assigns a constant correlation (a negligible negative value) to all pairs of units, and hence the optimized design performs similarly as the i.i.d. Gaussian design, as seen in Figure 8(b). It is because $f'(0)$ is almost zero in Figure 9(b), and thus the PGD-Gauss algorithm stops at the identity matrix, which is already a local optimizer.

A.5 Computational Aspect of PGD-Gauss

Here, we further discuss the computation cost of the proposed PGD-Gauss algorithm. We first present the computational complexity of PGD-Gauss in nuclear norm. From Algorithm 1, the computational complexity per iteration is

$$\underbrace{O(n^2(n+d))}_{\text{gradient descent}} + \underbrace{O(n^2)}_{\text{projection}} = O(n^2(n+d)).$$

As a first-order algorithm, PGD-Gauss exhibits local linear convergence with a proper initialization, and therefore it takes $O(\log(1/\epsilon))$ iterations to converge with numerical tolerance ϵ.

Then, we inspect the computational time in numerical studies. Under the setup in Section 7, we report average computational times for running PGD-Gauss over different experimental scales n (with a pre-specified number of iterations). From Table 4, we observe Time $\approx O(n^{2.12})$, which aligns with the theoretical computational complexity $O(n^2(n+d))$.

n	50	100	200	500
Time (s)	10.96	42.61	188.95	1324.03

Table 4. Average computation times for different sample sizes n.

B Main Proofs

Here we provide the core proofs related to Mehler's formula, asymptotic normality, and variance estimation. We also discuss inferential procedures under the continuous setting by the end of this section.

B.1 Mehler's Formula and Related Proofs

Here we prove results that are based on Mehler's formula, namely, Lemma 2 and Proposition 1. Proposition 2 is a direct result of Lemma 2 and its proof is omitted.

Proofs of Lemma 2. From Mehler's formula, for any $|\rho| \le 1$, it holds that

$$p_\rho(x,y) = \sum_{m=0}^{\infty} \rho^m h_m(x) h_m(y) \phi(x)\phi(y) \ .$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{X,Y} g(X)h(Y) &= \int g(x)h(y)p_\rho(x,y)\mathrm{d}x\mathrm{d}y \\
&= \sum_{m=0}^{\infty} \rho^m \int g(x)h(y)h_m(x)h_m(y)\phi(x)\phi(y)\mathrm{d}x\mathrm{d}y \\
&= \sum_{m=0}^{\infty} \rho^m \int g(x)h_m(x)\phi(x)\mathrm{d}x \int h(y)h_m(y)\phi(y)\mathrm{d}y \\
&= \sum_{m=0}^{\infty} \alpha_m[g]\alpha_m[h]\rho^m \ .
\end{aligned}$$

At the same time, by $h_0(x) = 1$ we notice

$$\mathbb{E}g(X) = \mathbb{E}g(X)h_0(X) = \alpha_0[g] \ , \quad \mathbb{E}h(X) = \mathbb{E}h(X)h_0(X) = \alpha_0[h] \ .$$

We have

$$\mathrm{Cov}_{X,Y}(g(X),h(Y)) = \sum_{m=0}^{\infty} \alpha_m[g]\alpha_m[h]\rho^m - \alpha_0[g]\alpha_0[h] = \sum_{m=1}^{\infty} \alpha_m[g]\alpha_m[h]\rho^m \ .$$

□

Proof of Proposition 1. By definition, for any $i \neq j$, the (i, j)-th entry of $\mathrm{Cov}_k(D)$ is

$$\mathrm{Cov}(\mathbb{I}\{D_i = k\}, \mathbb{I}\{D_j = k\}) \ .$$

Without loss of generality, we focus on $k = 2, \ldots, K-1$. The extreme cases $k = 1, K$ can be proved using a similar argument. Under the Gaussianization $D_i = g(T_i)$, we have

$$\begin{aligned}
\mathrm{Cov}(\mathbb{I}\{D_i = k\}, \mathbb{I}\{D_j = k\}) &= \mathrm{Cov}(\mathbb{I}\{T_i \in (q_{k-1}, q_k]\}, \mathbb{I}\{T_j \in (q_{k-1}, q_k]\}) \\
&= \mathrm{Cov}(\mathbb{I}\{T_i \leq q_k\} - \mathbb{I}\{T_i \leq q_{k-1}\}, \mathbb{I}\{T_j \leq q_k\} - \mathbb{I}\{T_j \leq q_{k-1}\}) \\
&= \mathrm{Cov}(\mathbb{I}\{T_i \leq q_k\}, \mathbb{I}\{T_j \leq q_k\}) + \mathrm{Cov}(\mathbb{I}\{T_i \leq q_{k-1}\}, \mathbb{I}\{T_j \leq q_{k-1}\}) \\
&\quad - 2\mathrm{Cov}(\mathbb{I}\{T_i \leq q_{k-1}\}, \mathbb{I}\{T_j \leq q_k\}) \\
&= r_{k,k}(\Sigma_{ij}) + r_{k-1.k-1}(\Sigma_{ij}) - 2r_{k,k-1}(\Sigma_{ij}) \ ,
\end{aligned}$$

where the last line follows by the definition of $r_{k,l}$ in Proposition 1.

Then it suffices to prove (6), i.e.,

$$\mathrm{Cov}(\mathbb{I}\{X \leq q_i\}, \mathbb{I}\{Y \leq q_j\}) = \int_0^\rho \frac{1}{2\pi\sqrt{1-r^2}} \exp\left(-\frac{q_i^2 + q_j^2 - 2rq_iq_j}{2(1-r^2)}\right) \mathrm{d}r \ .$$

Let $g(x) = \mathbb{I}\{x \leq q_i\}$ and $h(x) = \mathbb{I}\{x \leq q_j\}$. According to Lemma 2, for any $|\rho| \leq 1$, it holds that

$$r_{ij}(\rho) = \sum_{m=1}^{\infty} \alpha_m[g]\alpha_m[h]\rho^m \ .$$

For $\alpha_m[g]$, we derive that

$$\begin{aligned}
\alpha_m[g] &= \int_{-\infty}^{q_i} h_m(x)\phi(x)\mathrm{d}x \\
&= \frac{1}{\sqrt{m!}} \int_{-\infty}^{q_i} \mathrm{He}_m(x)\phi(x)\mathrm{d}x \\
&\overset{\text{(i)}}{=} \frac{1}{\sqrt{m!}} \int_{-\infty}^{q_i} (-1)^m \frac{\mathrm{d}^m}{\mathrm{d}x^m}\phi(x)\mathrm{d}x \\
&= \frac{-1}{\sqrt{m!}}\phi(x)\,\mathrm{He}_{m-1}(x)\Big|_{-\infty}^{q_i} \\
&\overset{\text{(ii)}}{=} -\frac{1}{\sqrt{m!}}\phi(q_i)\,\mathrm{He}_{m-1}(q_i) \ .
\end{aligned}$$

In the derivation above, (i) follows from the Definition 2, and (ii) follows from $\lim_{x\to-\infty} \phi(x)\,\mathrm{He}_m(x) = 0$. Hence, we have

$$\begin{aligned}
\alpha_m[g] &= -\frac{1}{\sqrt{m!}}\phi(q_i)\,\mathrm{He}_{m-1}(q_i) \ , \\
\alpha_m[h] &= -\frac{1}{\sqrt{m!}}\phi(q_j)\,\mathrm{He}_{m-1}(q_j) \ ,
\end{aligned}$$

where the proof of $\alpha_m[h]$ is identical. Based on Lemma 2, this implies

$$r_{ij}(\rho) = \sum_{m=1}^{\infty} \frac{1}{m!}\,\mathrm{He}_{m-1}(q_i)\,\mathrm{He}_{m-1}(q_i)\phi(q_i)\phi(q_j)\rho^m \ .$$

Notice that

$$r'_{ij}(\rho) = \sum_{m=1}^{\infty} \frac{1}{(m-1)!} \mathrm{He}_{m-1}(q_i)\,\mathrm{He}_{m-1}(q_i)\phi(q_i)\phi(q_j)\rho^{m-1} = p_\rho(q_i, q_j)\ ,$$
$$r_{ij}(0) = 0\ ,$$

where the first line follows from Mehler's formula. Then, by Newton–Leibniz theorem we obtain

$$r_{ij}(\rho) = \int_0^\rho p_r(q_i, q_j)\mathrm{d}r = \int_0^\rho \frac{1}{2\pi\sqrt{1-r^2}} \exp\left(-\frac{q_i^2 + q_j^2 - 2rq_iq_j}{2(1-r^2)}\right)\mathrm{d}r\ .$$

□

B.2 Asymptotics and Inference

In this section, we study the asymptotic properties of

$$\widehat{\tau}_k = \frac{K}{n}\sum_{i=1}^{n} Y_i\mathbb{I}\{D_i = k\}\ , \quad D_i = g(T_i)\ .$$

We prove its asymptotic normality in Theorem 1 and discuss the extensions. We defer the proof of Proposition 3 to Section C, as its proof follows a similar idea as those in supporting lemmas.

Our asymptotic analysis relies on Hájek's lemma [Lehmann and D'Abrera, 1975], which establishes the asymptotic equivalence between two sequences of random variables. We state below for completeness.

Lemma 3 (Hájek's Lemma). *If $(T_n - \mathbb{E}T_n)/\sqrt{\mathrm{Var}(T_n)}$ has a limit distribution $\mathcal{L}$ and if*

$$\frac{\mathbb{E}(T_n - S_n)^2}{\mathrm{Var}(T_n)} \to 0\ , \tag{15}$$

then $\mathrm{Var}(T_n)/\mathrm{Var}(S_n) \to 1$ and $(S_n - \mathbb{E}S_n)/\sqrt{\mathrm{Var}(S_n)}$ has the limit distribution $\mathcal{L}$.

In words, S_n and T_n share the same asymptotic distribution if the second moment of their difference is asymptotically smaller than $\mathrm{Var}(T_n)$.

B.2.1 Proof of Theorem 1. The crux of the proof is to establish an asymptotic equivalence between $\widehat{\tau}_k$ (hereafter denoted as $\widehat{\tau}^{opt}$) under $T \sim \mathcal{N}(0, \Sigma_\eta)$ and an ancillary estimator $\widehat{\tau}^{iid}$ under $T \sim \mathcal{N}(0, I_n)$. To this end, we proceed with the following steps.

(1) Construct a Hájek coupling $(\widehat{\tau}^{iid}, \widehat{\tau}^{opt})$.
(2) Establish the aforementioned asymptotic equivalence of $(\widehat{\tau}^{iid}, \widehat{\tau}^{opt})$ using Hájek's lemma.
(3) Prove the asymptotic normality for $\widehat{\tau}^{iid}$.

Step 1. Construct Hájek's coupling. To construct $(\widehat{\tau}^{iid}, \widehat{\tau}^{opt})$, we first define

$$T^{iid} \sim \mathcal{N}(0, I_n)\ , \quad T^{opt} = \Sigma_\eta^{1/2} T^{iid}\ .$$

One can easily check that $T^{opt} \sim \mathcal{N}(0, \Sigma_\eta)$ and $\mathrm{Cov}(T^{iid}, T^{opt}) = \Sigma_\eta^{1/2}$. Then, define

$$\widehat{\tau}^{iid} = \frac{K}{n}\sum_{i=1}^{n} \mathbb{I}\{g(T_i^{iid}) = k\}\tilde{Y}_i(k) + \frac{1}{n}\sum_{i=1}^{n}\left(Y_i(k) - \widetilde{Y}_i(k)\right)\ ,$$
$$\widehat{\tau}^{opt} = \frac{K}{n}\sum_{i=1}^{n} \mathbb{I}\{g(T_i^{opt}) = k\}Y_i(k)\ .$$

$\widehat{\tau}^{iid}$ matches the distribution of $\widehat{\tau}^{opt}$, since

$$\mathbb{E}\widehat{\tau}^{iid} = \mathbb{E}\widehat{\tau}^{opt} = \tau_k \ .$$

More importantly, their variances also match, since

$$\begin{aligned}
\mathrm{Var}(\widehat{\tau}^{iid}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} K\mathbb{I}\{g(T_i^{iid}) = k\}\tilde{Y}_i(k)\right) \\
&\overset{\text{(i)}}{=} \frac{K^2}{n^2}\tilde{Y}(k)^\top \mathrm{Cov}(D_k^{iid})\tilde{Y}(k) \\
&\overset{\text{(ii)}}{=} \frac{K^2}{n^2}\tilde{Y}(k)^\top f_k(I_n)\tilde{Y}(k) \\
&\overset{\text{(iii)}}{=} \frac{K^2}{n^2}Y(k)^\top f_k(\Sigma_\eta)^{1/2} f_k(I_n)^{-1/2} f_k(I_n) f_k(I_n)^{-1/2} f_k(\Sigma_\eta) Y(k) \\
&= \frac{K^2}{n^2}Y(k)^\top f_k(\Sigma_\eta) Y(k) = \mathrm{Var}(\widehat{\tau}^{opt}) \ .
\end{aligned}$$

In (i), D_k^{iid} denotes the treatment vector $(\mathbb{I}\{g(T_1^{iid}) = k\}, \ldots, \mathbb{I}\{g(T_n^{iid}) = k\})$; (ii) follows from Mehler's formula and Proposition 1; (iii) follows from the definition of $\tilde{Y}(k)$.

Step 2. Establish asymptotic equivalence. Based on Hájek's Lemma (Lemma 3), we need to verify (15) for $(\widehat{\tau}^{iid}, \widehat{\tau}^{opt})$, that is

$$\frac{\mathbb{E}(\widehat{\tau}^{iid} - \widehat{\tau}^{opt})^2}{\mathrm{Var}(\widehat{\tau}^{iid})} \to 0 \ .$$

Observe that

$$\mathrm{Var}(\widehat{\tau}^{iid}) = \frac{K-1}{n^2}\sum_{i=1}^{n}\tilde{Y}_i(k)^2 \ .$$

Under the assumptions in Theorem 1, $n\,\mathrm{Var}(\widehat{\tau}^{iid})$ converges to a positive limit, and thus $\mathrm{Var}(\widehat{\tau}^{iid}) \asymp 1/n$. Then, it suffices to verify

$$n\mathbb{E}(\widehat{\tau}^{iid} - \widehat{\tau}^{opt})^2 \to 0 \ . \tag{16}$$

Notice that

$$\begin{aligned}
\mathbb{E}(\widehat{\tau}^{iid} - \widehat{\tau}^{opt})^2 &\overset{\text{(i)}}{=} \mathrm{Var}(\widehat{\tau}^{iid} - \widehat{\tau}^{opt}) \\
&= \frac{K^2}{n^2}\,\mathrm{Var}\left(\sum_{i=1}^{n}(\mathbb{I}\{g(T_i^{iid}) = k\}\tilde{Y}_i(k) - \mathbb{I}\{g(T_i^{opt}) = k\}Y_i(k))\right) \\
&= \frac{K^2}{n^2}\left(\tilde{Y}(k)^\top \mathrm{Cov}(D_k^{iid})\tilde{Y}(k) - 2\tilde{Y}(k)^\top \mathrm{Cov}(D_k^{iid}, D_k^{opt})Y(k) + Y(k)^\top \mathrm{Cov}(D_k^{opt})Y(k)\right) \\
&\overset{\text{(ii)}}{=} \frac{K^2}{n^2}\left(\tilde{Y}(k)^\top f_k(I_n)\tilde{Y}(k) - 2\tilde{Y}(k)^\top f_k(\Sigma_\eta^{1/2})Y(k) + Y(k)^\top f_k(\Sigma_\eta)Y(k)\right) ,
\end{aligned}$$

where (i) follows from $\mathbb{E}\widehat{\tau}^{iid} = \mathbb{E}\widehat{\tau}^{opt}$, and (ii) follows from Proposition 1.

By Proposition 1, it is easy to verify that $f_k(0) = 0$ and $f_k(1) = (K-1)/K^2$. From now on, without loss of generality, we may rescale f_k such that $f_k(0) = 0$ and $f_k(1) = 1$. This does not affect the order of the quantity above, since K is a fixed constant. After rescaling, we have $f_k(I_n) = I_n$, which

simplifies the derivation below. By definition of $\tilde{Y}(k)$, we have

$$\begin{aligned}\mathbb{E}(\widehat{\tau}^{iid} - \widehat{\tau}^{opt})^2 &= \frac{K^2}{n^2} Y(k)^\top \left(2f_k(\Sigma_\eta) - 2f_k(\Sigma_\eta)^{1/2} f_k(\Sigma_\eta^{1/2})\right) Y(k)^\top \,, \\ &\le \frac{2K^2}{n^2} \|f_k(\Sigma_\eta) - f_k(\Sigma_\eta)^{1/2} f_k(\Sigma_\eta^{1/2})\|_{\mathrm{op}} \|Y(k)\|^2 \\ &\overset{\text{(i)}}{\le} \frac{2MK^2}{n} \|f_k(\Sigma_\eta)\|_{\mathrm{op}}^{1/2} \|f_k(\Sigma_\eta)^{1/2} - f_k(\Sigma_\eta^{1/2})\|_{\mathrm{op}} \,, \end{aligned} \tag{17}$$

where (i) follows from $\|Y(k)\|^2 \le nM$.

To analyze the operator norm above, we first give a decomposition of Σ_η.

Lemma 4. *Suppose Assumption 1 holds. In the one-step PGD-Gauss, the obtained solution Σ_η satisfies a decomposition*

$$\Sigma_\eta = I_n + \eta N \,,$$

where N is a symmetric matrix with zero diagonal values, and

$$\|N\|_{\mathrm{op}} = O(\|XX^\top - I_n\|_{\mathrm{op}} + \eta \|XX^\top - I_n\|_{\mathrm{op}}^2) \,.$$

Next we introduce the following result based on Taylor expansions.

Lemma 5. *For $\Sigma \in \mathcal{E}$, define $\Delta = \Sigma - I_n$, the residual matrix with zero diagonal values. Suppose $\|\Delta\|_{\mathrm{op}} = o(1)$. We have*

$$\begin{aligned} f_k(\Sigma) &= I_n + f_k'(0)\Delta + R_1 \,, \quad \|R_1\|_{\mathrm{op}} = O(\|\Delta\|_{\mathrm{op}}^2) \,, \quad \|f_k(\Sigma)\|_{\mathrm{op}} = O(1) \\ f_k(\Sigma)^{1/2} &= I_n + \frac{1}{2} f_k'(0)\Delta + R_2 \,, \quad \|R_2\|_{\mathrm{op}} = O(\|\Delta\|_{\mathrm{op}}^2) \,, \\ f_k(\Sigma^{1/2}) &= I_n + \frac{1}{2} f_k'(0)\Delta + R_3 \,, \quad \|R_3\|_{\mathrm{op}} = o(1) + O(\|\Delta\|_{\mathrm{op}}^2) \,. \end{aligned}$$

Moreover, the operator norm of R_1, R_2, R_3 are all of order $o(1)$ since $\|\Delta\|_{\mathrm{op}} = o(1)$.

Now, we utilize Lemma 4 and Lemma 5 to verify the Hájek condition (16). Based on Lemma 4, we have

$$\begin{aligned} \Sigma_\eta = I_n + \Delta \,, \quad \Delta = \eta N \,, \\ \|N\|_{\mathrm{op}} = O(\|XX^\top - I_n\|_{\mathrm{op}} + \eta \|XX^\top - I_n\|_{\mathrm{op}}^2) \,, \\ \Rightarrow \|\Delta\|_{\mathrm{op}} = O(\eta \|XX^\top - I_n\|_{\mathrm{op}} + \eta^2 \|XX^\top - I_n\|_{\mathrm{op}}^2) \,, \end{aligned}$$

where N is a symmetric matrix with zero diagonal values. Under Assumption 1 that $\eta\|XX^\top - I_n\|_{\mathrm{op}} = o(1)$, one can verify that

$$\|\Delta\|_{\mathrm{op}} = o(1) \,.$$

Thus the condition required in Lemma 5 is satisfied. We apply Lemma 5 to obtain

$$\begin{aligned} \|f_k(\Sigma_\eta)\|_{\mathrm{op}} &= O(1) \,, \\ f_k(\Sigma_\eta)^{1/2} &= I_n + \frac{1}{2} f_k'(0)\eta N + R_2 \,, \quad \|R_2\|_{\mathrm{op}} = o(1) \,, \\ f_k(\Sigma_\eta^{1/2}) &= I_n + \frac{1}{2} f_k'(0)\eta N + R_3 \,, \quad \|R_3\|_{\mathrm{op}} = o(1) \,. \end{aligned} \tag{18}$$

By applying Equations (18) to (17), we obtain

$$\mathbb{E}(\widehat{\tau}^{iid} - \widehat{\tau}^{opt})^2 = O\left(\frac{1}{n} \|f_k(\Sigma_\eta)\|_{\mathrm{op}}^{1/2} \|f_k(\Sigma_\eta)^{1/2} - f_k(\Sigma_\eta^{1/2})\|_{\mathrm{op}}\right) = o\left(\frac{1}{n}\right) \,.$$

Therefore, condition (16) holds and one can apply Hájek's Lemma (Lemma 3) to obtain that

$$\frac{\widehat{\tau}^{opt} - \tau_k}{\sqrt{\mathrm{Var}(\widehat{\tau}^{opt})}}$$

has the same asymptotic distribution as

$$\frac{\widehat{\tau}^{iid} - \tau_k}{\sqrt{\mathrm{Var}(\widehat{\tau}^{iid})}} .$$

Step 3. Asymptotic normality for $\widehat{\tau}^{iid}$. We define

$$X_{ni} = (K\mathbb{I}\{g(T_i^{iid}) = k\} - 1)\tilde{Y}_i(k) , \quad S_n = \sum_{i=1}^{n} X_{ni} .$$

It is then easy to verify that

$$\widehat{\tau}^{iid} - \tau_k = \frac{1}{n} S_n , \quad \mathrm{Var}(\widehat{\tau}^{iid}) = \mathrm{Var}(\frac{1}{n} S_n) .$$

Therefore, it suffices to derive the asymptotic distribution for S_n. The Lindeberg condition requires that for any $\varepsilon > 0$,

$$\frac{1}{\mathrm{Var}(S_n)} \sum_{i=1}^{n} \mathbb{E} X_{ni}^2 \mathbb{I}\{X_{ni}^2 \geq \varepsilon \, \mathrm{Var}(S_n)\} \to 0 .$$

Note that

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n} \mathrm{Var}(X_{ni}) = (K-1) \sum_{i=1}^{n} \tilde{Y}_i^2(k) .$$

Under Condition 2 in Theorem 1, $\mathrm{Var}(S_n)$ is of order n. Hence,

$$\frac{\max_i X_{ni}^2}{\mathrm{Var}(S_n)} \leq \frac{\max_i (K-1)\tilde{Y}_i^2(k)}{\mathrm{Var}(S_n)} \asymp \frac{\max_i \tilde{Y}_i^2(k)}{n} = o(1) .$$

The last equality follows from Condition 1. This suggests that all the summands in the Lindeberg condition become zero for large n. Therefore, the Lindeberg condition is satisfied, and we have

$$\frac{S_n}{\sqrt{\mathrm{Var}(S_n)}} \xrightarrow{d} \mathcal{N}(0,1) .$$

Note that $\mathrm{Var}(S_n) = (K-1) \sum Y_i^2(k)$. We have

$$\sqrt{n}(\widehat{\tau}^{iid} - \tau_k) \xrightarrow{d} \mathcal{N}(0, \lim_n \frac{K-1}{n} \sum_i \tilde{Y}_i^2(k)) .$$

This completes the proof of Theorem 1.

B.2.2 Generalization to Multi-Step PGD-Gauss Solutions. Here, we discuss a generalization of Theorem 1 to multi-step PGD-Gauss solutions. Specifically, the proof of Theorem 1 indicates a more general result below.

Corollary 1. *Consider a Gaussianization $T \sim \mathcal{N}(0, \Sigma)$. Suppose that $\Delta \coloneqq \Sigma - I_n$ satisfies $\|\Delta\|_{\mathrm{op}} = o(1)$, and that Conditions 1-3 in Theorem 1 hold. Then, we have*

$$\sqrt{n}\,(\widehat{\tau}_k - \tau_k) \xrightarrow{d} \mathcal{N}\left(0, \lim_{n\to\infty} \frac{K-1}{n} \|\tilde{Y}(k)\|^2\right) .$$

Corollary 1 can be viewed as a result for general PGD-Gauss solutions. Regardless of how many steps taken in the PGD-Gauss, the asymptotic normality for $\widehat{\tau}_k$ holds as long as the solution Σ does not deviate too much from the identity matrix.

Proof. Following Step 1 in the proof of Theorem 1, we construct the coupling in the same way. Then, based on the analysis in Step 2, it suffices to show that

$$\|f_k(\Sigma)\|_{\mathrm{op}}^{1/2}\|f_k(\Sigma)^{1/2} - f_k(\Sigma^{1/2})\|_{\mathrm{op}} = o(1)\ , \quad \Sigma = I_n + \Delta\ .$$

Under our assumption in Corollary 1, we have $\|\Delta\|_{op} = o(1)$ and thus the condition in Lemma 5 is satisfied. We can then apply Lemma 5 to show the above equation. Lastly, we can apply Step 3 in Theorem 1 and complete the proof. □

B.2.3 Inference. First, we prove Theorem 2 by showing that the variance of $\widehat{V}_{iid}$ converges to zero.

Proof of Theorem 2. Define $Y(k)^2 = (Y_1(k)^2, \ldots, Y_n(k)^2) \in \mathbb{R}^n$. Then, we can write

$$\widehat{V}_{iid} = \frac{(K-1)K}{n}\sum_i Y_i(k)^2\mathbb{I}\{D_i = k\}\ , \quad D_i = g(T_i)\ .$$

By the variance formula of Horvitz-Thompson estimator (Lemma 1), we have

$$\mathrm{Var}(\widehat{V}_{iid}) = \frac{(K-1)^2K^2}{n^2}Y(k)^{2\top}\mathrm{Cov}_k(D)Y(k)^2\ .$$

Under i.i.d. Gaussianization, one can show that $\mathrm{Cov}_k(D) = (K-1)I_n/K^2$, i.e., a rescaled identity matrix. Hence,

$$\mathrm{Var}(\widehat{V}_{iid}) = \frac{(K-1)^3}{n^2}\sum_{i=1}^n Y_i(k)^2 = O\left(\frac{1}{n}\right)\ .$$

The last equality follows from the fact that $\max_i |Y_i(k)| = O(1)$ and K is constant. □

Next we present general variance estimators under our Gaussianization framework. Concretely, we focus on the following estimator defined in Section 5:

$$\widehat{\tau}_w^c = \frac{1}{n}\sum_{i=1}^n Y_i(T_i)w(T_i) = \frac{1}{n}\sum Y_iW_i\ , \quad W_i \coloneqq w(T_i)\ .$$

Note that by setting $w(t) = K\sum_{k=1}^K w_k\mathbb{I}\{g(t) = k\}$, we recover the general estimator $\widehat{\tau}_w$ in Section 3. Therefore, the estimator above applies to both discrete and continuous treatments.

The variance of $\widehat{\tau}_w^c$ is

$$V_{iid} = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(Y_iW_i)\ .$$

We estimate the variance of $\widehat{\tau}_w^c$ by the sample variance of $\{Y_iW_i\}_{i=1}^n$, that is,

$$\widehat{V}_{iid} = \frac{1}{n-1}\sum_{i=1}^n\left(Y_iW_i - \frac{1}{n}\sum_{i=1}^n Y_iW_i\right)^2\ . \tag{19}$$

Compared to Theorem 2, $\widehat{V}_{iid}$ in (19) is no longer unbiased, but it is a valid variance estimator with a nonnegative bias.

Lemma 6. *Under i.i.d. Gaussianization $T \sim \mathcal{N}(0, I_n)$ with the variance estimator* (19)*, we have*

$$\mathbb{E}\widehat{V}_{iid} = V_{iid} + \frac{1}{n-1}\sum_{i=1}^n\left(\mathbb{E}Y_iW_i - \frac{1}{n}\sum_{i=1}^n\mathbb{E}Y_iW_i\right)^2 \geq V_{iid}\ .$$

Proof. For notational simplicity, we define $R_i = Y_i W_i$, such that $\widehat{V}_{iid}$ can be written as

$$\widehat{V}_{iid} = \frac{1}{n-1}\sum_{i=1}^{n}\left(R_i - \frac{1}{n}\sum_{i=1}^{n} R_i\right)^2 .$$

Then, we have

$$\widehat{V}_{iid} = \frac{1}{n-1}\sum_{i=1}^{n} R_i^2 - \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} R_i\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n} R_i^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} R_i^2 + \sum_{i\neq j} R_i R_j\right),$$

$$\mathbb{E}\widehat{V}_{iid} = \frac{1}{n-1}\sum_{i=1}^{n} \mathbb{E}R_i^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} \mathbb{E}R_i^2 + \sum_{i\neq j} \mathbb{E}R_i \mathbb{E}R_j\right).$$

In the second line, we use the fact that $R_i \perp\!\!\!\perp R_j$ since $T_i \perp\!\!\!\perp T_j$. Next,

$$\begin{aligned}
\mathbb{E}\widehat{V}_{iid} &= \frac{1}{n-1}\sum_{i=1}^{n} \mathbb{E}R_i^2 - \frac{1}{n(n-1)}\left(\sum_{i=1}^{n} \mathbb{E}R_i^2 + \sum_{i\neq j} \mathbb{E}R_i \mathbb{E}R_j + (n-1)\sum_{i=1}^{n}(\mathbb{E}R_i)^2 - (n-1)\sum_{i=1}^{n}(\mathbb{E}R_i)^2\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n} \mathbb{E}R_i^2 - \frac{1}{n(n-1)}\sum_{i=1}^{n} \mathbb{E}R_i^2 - \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}R_i)^2 + \frac{1}{n(n-1)}\left((n-1)\sum_{i=1}^{n}(\mathbb{E}R_i)^2 - \sum_{i\neq j} \mathbb{E}R_i \mathbb{E}R_j\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}R_i^2 - \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}R_i)^2 + \frac{1}{n(n-1)}\left(n\sum_{i=1}^{n}(\mathbb{E}R_i)^2 - \sum_{i,j=1}^{n} \mathbb{E}R_i \mathbb{E}R_j\right) \\
&= V_{iid} + \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathbb{E}R_i - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}R_i\right)^2 .
\end{aligned}$$

Lastly, note that $\mathbb{E}R_i = Y_i W_i$. This gives the mean formula of $\widehat{V}_{iid}$. □

Based on Lemma 6, we construct normality-based confidence intervals by

$$\left[\widehat{\tau}_w^c - z_{\alpha/2}\sqrt{\frac{\widehat{V}_{iid}^c}{n}},\ \widehat{\tau}_w^c + z_{\alpha/2}\sqrt{\frac{\widehat{V}_{iid}^c}{n}}\right].$$

Asymptotic validity of this approach can be established similarly as in Section 6.

For $T \sim \mathcal{N}(0, \Sigma)$ with a general covariance matrix, we again employ randomization-based inference as described below.

Procedure 3 (Randomization-Based Confidence Interval for $\widehat{\tau}_w^c$).

(1) *Fit a model $\widehat{m}(X_i, T_i)$ by regressing Y_i over X_i and T_i, $i = 1, \ldots, n$.*

(2) *Generate $\{T^b\}_{b=1}^{B} \overset{iid}{\sim} \mathcal{N}(0, \Sigma)$. For each randomization T^b, impute outcomes by*

$$Y_i^b = \begin{cases} Y_i & \text{if } T_i^b = T_i \\ \widehat{m}(X_i, T_i^b) & \text{if } T_i^b \neq T_i \end{cases}.$$

Compute the randomization-based estimate $\widehat{\tau}_w^{c,b}$ based on T_i^b and Y_i^b, $i = 1, \ldots, n$.

(3) *Construct the randomization-based confidence interval*

$$[\widehat{c}(\alpha/2),\ \widehat{c}(1-\alpha/2)] .$$

Here, $\widehat{c}(\alpha)$ is the α-sample quantile for $\{\widehat{\tau}_k^{c,b}\}_{b=1}^{B}$.

C Proofs of Supporting Lemmas

In the proof of supporting lemmas, we utilize matrix norm inequalities to analyze the perturbation of a Gaussian covariance Σ with respect to I_n. Specifically, for any matrix A, we have

$$\|A\|_{\mathrm{op}} \le \sqrt{\|A\|_1 \|A\|_\infty} \,. \tag{20}$$

where $\|A\|_1$ and $\|A\|_\infty$ denote the matrix 1-norm and infinity-norm, i.e.,

$$\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{n} |A_{ij}| \,, \quad \|A\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^{n} |A_{ij}| \,.$$

Moreover, when A is a symmetric matrix, we have $\|A\|_1 = \|A\|_\infty$, and

$$\|A\|_{\mathrm{op}} \le \|A\|_\infty \,. \tag{21}$$

These inequalities will be invoked multiple times in our proof. Additionally, for $A \in \mathbb{R}^{n\times n}$, let $\mathrm{diag}(A) \in \mathbb{R}^{n\times n}$ be the diagonal matrix with the same diagonal values as A.

Our proof leverages the specific form of the one-step PGD-Gauss under the nuclear norm. That is, based on Algorithm 1, we have

$$\begin{aligned} \Sigma_\eta &= V_\eta V_\eta^\top \,, \quad V_\eta = D^{-1} U_\eta \,, \\ U_\eta &= (I_n - \eta \nabla l_{\mathrm{norm}}(I_n)) V_0 = I_n - \eta \nabla l_{\mathrm{norm}}(I_n) \,, \\ \nabla l_{\mathrm{norm}}(I_n) &= f_k'(0)(XX^\top - I_n) \,, \end{aligned} \tag{22}$$

where D is a diagonal matrix with i-th diagonal equal to the norm of u_i, the i-th row of U_η.

Lastly, we prove the following matrix inequality.

Lemma 7. *For any symmetric matrix A, we have* $\max_{ij} |A_{ij}| \le 2\|A\|_{\mathrm{op}}$.

Proof. By the variational expression of the operator norm, we have

$$\|A\|_{\mathrm{op}} = \sup_{\|x\|=1} |x^\top A x| \,.$$

Let $x = e_i$, the basis vector with the i-th entry equal to one, we obtain

$$|x^\top A x| = |A_{ii}| \le \|A\|_{\mathrm{op}} \,.$$

Then, for any $i \neq j$, by setting $x = (e_i + e_j)/\sqrt{2}$, we obtain

$$|x^\top A x| = \frac{1}{2} |A_{ii} + A_{jj} + 2A_{ij}| \le \|A\|_{\mathrm{op}} \,.$$

By the triangular inequality, we have $|A_{ii} + A_{jj} + 2A_{ij}| \ge \big||A_{ii} + A_{jj}| - 2|A_{ij}|\big|$, and hence

$$\big||A_{ii} + A_{jj}| - 2|A_{ij}|\big| \le 2\|A\|_{\mathrm{op}} \,.$$

This implies

$$2|A_{ij}| \le 2\|A\|_{\mathrm{op}} + |A_{ii} + A_{jj}| \le 4\|A\|_{\mathrm{op}} \,.$$

Therefore, we obtain $\max_{ij} |A_{ij}| \le 2\|A\|_{\mathrm{op}}$. □

C.1 Proof of Lemma 4

Based on (22), we have

$$D_{ii} = \sqrt{1 + \sum_{j \neq i} \eta^2 \nabla^2 l_{\text{norm},ij}(I_n)} \ .$$

Notice that $\sum_{j\neq i} \nabla^2 l_{\text{norm},ij}(I_n)$ is the i-th diagonal of matrix $(\nabla l_{\text{norm}}(I_n))^2$, we have

$$|\sum_{j\neq i} \nabla^2 l_{\text{norm},ij}(I_n)| \leq \|\nabla l_{\text{norm}}(I_n)\|_{\text{op}}^2 = (f_k'(0))^2 \|XX^\top - I_n\|_{\text{op}}^2 \ .$$

Thus,

$$D_{ii} \leq \sqrt{1 + \eta^2 \|XX^\top - I_n\|_{\text{op}}^2 (f_k'(0))^2} = \sqrt{1 + O(\eta^2 \|XX^\top - I_n\|_{\text{op}}^2)} \ . \tag{23}$$

Under Assumption 1, $D_{ii} = \sqrt{1 + o(1)}$. Therefore, $\max |D_{ii} - 1| = o(1)$. This fact will be used in our proof.

Noticing that $\Sigma_\eta = V_\eta V_\eta^\top$, we have

$$\begin{aligned}
\Sigma_\eta &= D^{-1}(I_n - \eta \nabla l_{\text{norm}}(I_n))^2 D^{-1} \\
&= D^{-1}(I_n - \eta f_k'(0)(XX^\top - I_n))^2 D^{-1} \\
&= D^{-1}(I_n - 2\eta f_k'(0)(XX^\top - I_n) + \eta^2 (f_k'(0)(XX^\top - I_n))^2) D^{-1} \\
&= D^{-2} + \underbrace{D^{-1}(-2\eta f_k'(0)(XX^\top - I_n) + \eta^2 (f_k'(0)(XX^\top - I_n))^2) D^{-1}}_{=: M} \ ,
\end{aligned} \tag{24}$$

where we use M to denote the component that contributes to off-diagonal elements.

We write

$$\Sigma_\eta = I_n + \eta N \ , \quad N \coloneqq \frac{1}{\eta}(\Sigma_\eta - I_n) \ .$$

Based on the M defined in Equation (24), we derive

$$N = \frac{1}{\eta}(D^{-2} + M - I_n) = \frac{1}{\eta}(M - \text{diag}(M)) \ . \tag{25}$$

Hence,

$$\eta \|N\|_{\text{op}} \leq \|M\|_{\text{op}} + \|\text{diag}(M)\|_{\text{op}} \ .$$

For $\|M\|_{\text{op}}$, we have

$$\begin{aligned}
\|M\|_{\text{op}} &\leq \|D^{-1}\|_{\text{op}}^2 \left(2\eta \|f_k'(0)(XX^\top - I_n)\|_{\text{op}} + \eta^2 \|f_k'(0)(XX^\top - I_n)\|_{\text{op}}^2\right) \\
&= O\left(\eta \|XX^\top - I_n\|_{\text{op}} + \eta^2 \|XX^\top - I_n\|_{\text{op}}^2\right) \ .
\end{aligned}$$

The last line follows from $\|D^{-1}\|_{\text{op}} = 1 + o(1)$ since $\max |D_{ii} - 1| = o(1)$. Similarly, for $\|\text{diag}(M)\|_{\text{op}}$, we have

$$\begin{aligned}
\|\text{diag}(M)\|_{\text{op}} &= O\left(\eta \|\text{diag}(f'(0)(XX^\top - I_n))\|_{\text{op}}\right) + O\left(\eta^2 \|\text{diag}((f'(0)(XX^\top - I_n))^2)\|_{\text{op}}\right) \\
&\overset{\text{(i)}}{=} O\left(\eta^2 \|\text{diag}((XX^\top - I_n)^2)\|_{\text{op}}\right) \overset{\text{(ii)}}{=} O\left(\eta^2 \|XX^\top - I_n)\|_{\text{op}}^2\right) \ .
\end{aligned}$$

In (i), we use the fact that $\text{diag}(XX^\top - I_n) = 0$, since $\|X_i\| = 1$ under Assumption 1. (ii) follows from the fact that $\|\text{diag}(A)\|_{\text{op}} \leq \|A\|_{\text{op}}$ for a positive semidefinite matrix A.

Based on our analysis for $\|M\|_{\text{op}}$ and $\|\text{diag}(M)\|_{\text{op}}$ above, we have

$$\|N\|_{\text{op}} = O\left(\|XX^\top - I_n\|_{\text{op}} + \eta \|XX^\top - I_n\|_{\text{op}}^2\right) \ .$$

C.2 Proof of Lemma 5

Here we prove the results in Lemma 5 one by one. Recall that we assume $f_k(1) = 1$, such that $f_k(I_n) = I_n$, as explained in the main proof of Section B.2.

Analyze $f_k(\Sigma)$. Note that f_k is smooth around zero. For any $x \in (-1, 1)$, we use Taylor expansion to obtain

$$f_k(x) = f_k(0) + f_k'(0)x + \frac{1}{2}f_k''(\xi_x)x^2 = f_k'(0)x + \frac{1}{2}f_k''(\xi_x)x^2 \,, \tag{26}$$

where ξ_x is a constant satisfying $|\xi_x| \le |x|$. With the matrix input Σ, we apply the above Taylor expansion to off-diagonal entries to obtain

$$f_k(\Sigma) = I_n + f_k'(0)\Delta + R_1 \,, \quad R_{1,ij} = \frac{1}{2}f_k''(\xi_{ij})\Delta_{ij}^2 \,,$$

where ξ_{ij} satisfies $|\xi_{ij}| \le |\Delta_{ij}|$. Since R_1 is symmetric, by (21), we have $\|R_1\|_{\mathrm{op}} \le \|R_1\|_\infty$ and

$$\|R_1\|_\infty = \max_{i=1,\dots,n} \sum_{j=1}^n |R_{f,ij}| = \frac{1}{2}\max_i \sum_j |f_k''(\xi_{ij})\Delta_{ij}^2| \,.$$

By Lemma 7, $\max_{ij}|\Delta_{ij}| = O(\|\Delta\|_{\mathrm{op}}) = o(1)$. Since $|\xi_{ij}| \le |\Delta_{ij}|$, we have $\max_{ij}|\xi_{ij}| \le \max_{ij}|\Delta_{ij}| = o(1)$, and hence $\max_{ij}|f_k''(\xi_{ij})| = O(1)$. We have

$$\sum_i |f_k''(\xi_{ij})\Delta_{ij}^2| = O(\sum_i \Delta_{ij}^2) \,.$$

Notice that $\sum_i \Delta_{ij}^2$ is the j-th diagonal of the squared matrix Δ^2. Therefore, we have

$$O(\sum_i \Delta_{ij}^2) = O(\|\Delta^2\|_{\mathrm{op}}) = O(\|\Delta\|_{\mathrm{op}}^2) \,.$$

To sum up, we derive $\|R_1\|_\infty = O(\|\Delta\|_{\mathrm{op}}^2)$, and thus $\|R_1\|_{\mathrm{op}} = O(\|\Delta\|_{\mathrm{op}}^2)$. In addition, under the condition $\|\Delta\|_{\mathrm{op}} = o(1)$, we have

$$\|f_k(\Sigma)\|_{\mathrm{op}} \le \|I_n\|_{\mathrm{op}} + |f_k'(0)|\|\Delta\|_{\mathrm{op}} + \|R_1\|_{\mathrm{op}} = O(1 + \|\Delta\|_{\mathrm{op}} + \|\Delta\|_{\mathrm{op}}^2) = O(1) \,.$$

Analyze $f_k(\Sigma)^{1/2}$. We first introduce the Taylor expansion of the matrix square root. For any symmetric matrix with $\|M\|_{\mathrm{op}} < 1$, we define its eigenvalue decomposition as $M = U\Lambda U^\top$. Then, by definition of the matrix square root,

$$(I_n + M)^{1/2} = U(I_n + \Lambda)^{1/2}U^\top \,.$$

That is, the i-th eigenvalue $(1 + \Lambda_i)$ is transformed to $s(1 + \Lambda_i)$, where $s(x) = \sqrt{x}$ is the square-root function. Based on Taylor expansion of $s(\cdot)$,

$$s(1 + \Lambda_i) = 1 + \frac{1}{2}\Lambda_i + \frac{1}{2}s''(\xi_i)\Lambda_i^2 \,,$$

where ξ_i is some value satisfying $|\xi_i - 1| \le |\Lambda_i| < 1$. Therefore, we can write

$$(I_n + M)^{1/2} = U\left(I_n + \frac{1}{2}\Lambda + \frac{1}{2}S\Lambda^2\right)U^\top = I_n + \frac{1}{2}M + \frac{1}{2}US\Lambda^2U^\top \,,$$

where S is a diagonal matrix with elements $s''(\xi_i)$.

For $f_k(\Sigma)$, we may set

$$\Delta_f = f_k(\Sigma) - I_n = f_k'(0)\Delta + R_1 \,.$$

Moreover, $\|\Delta_f\|_{\mathrm{op}} \le |f_k'(0)|\|\Delta\|_{\mathrm{op}} + \|R_1\|_{\mathrm{op}} = O(\|\Delta\|_{\mathrm{op}}) + O(\|\Delta\|_{\mathrm{op}}^2) = o(1)$. Therefore, we can apply the matrix square root Taylor expansion to obtain

$$\begin{aligned} f_k(\Sigma)^{1/2} &= I_n + \frac{1}{2}\Delta_f + \frac{1}{2}US\Lambda^2U^\top \\ &= I_n + \frac{1}{2}f_k'(0)\Delta + \frac{1}{2}R_1 + R_f \ , \\ R_f &\coloneqq \frac{1}{2}US\Lambda^2U^\top \ . \end{aligned}$$

Here, the matrices U, S, Λ are defined with respect to $M = \Delta_f$. Then, it remains to bound the operator norm of R_f. Note that $\|\Delta_f\|_{\mathrm{op}} = o(1)$ and $s(\cdot)$ is smooth around one. We have $\max_i\{|s''(\xi_i)|\} = O(1)$. Therefore, the norm of R_f can be bounded as

$$\|R_f\|_{\mathrm{op}} \le O(\|\Delta_f\|_{\mathrm{op}}^2) = O(\|\Delta\|_{\mathrm{op}}^2 + \|\Delta\|_{\mathrm{op}}^4) = O(\|\Delta\|_{\mathrm{op}}^2) \ .$$

By setting $R_2 = R_1/2 + R_f$, we obtain

$$f(\Sigma)^{1/2} = I_n + \frac{1}{2}f_k'(0)\Delta + R_2 \ , \quad \|R_2\|_{\mathrm{op}} = O(\|\Delta\|_{\mathrm{op}}^2) \ .$$

Analyze $f_k(\Sigma^{1/2})$. First, we apply the matrix square root Taylor expansion to $\Sigma = I_n + \Delta$ to obtain

$$\begin{aligned} \Sigma^{1/2} &= I_n + \frac{1}{2}\Delta + R_s \ , \\ R_s &\coloneqq \frac{1}{2}US\Lambda^2U^\top \ , \end{aligned}$$

where U, S, Λ are defined with respect to $M = \Delta$. Using the same logic as in the last step, we can show $\max_i\{|s''(\xi_i)|\} = O(1)$. Therefore, R_s can be bounded as

$$\|R_s\|_{\mathrm{op}} = O(\|\Delta\|_{\mathrm{op}}^2) = o(1) \ . \tag{27}$$

Moreover, Lemma 7 indicates that

$$\max_{ij}|R_{s,ij}| = O(\|R_s\|_{\mathrm{op}}) = o(1) \ . \tag{28}$$

To analyze $f_k(\Sigma^{1/2})$, we observe that

$$\begin{aligned} f_k(\Sigma^{1/2}) &= f_k(I_n + \frac{1}{2}\Delta + R_s) \\ &= f_k(I_n + \mathrm{diag}(R_s)) + f_k(\frac{1}{2}\Delta + R_s - \mathrm{diag}(R_s)) \ , \end{aligned}$$

In the second line, we decompose the matrix into a diagonal matrix $f_k(I_n + \mathrm{diag}(R_s))$ and an off-diagonal matrix $f_k(\frac{1}{2}\Delta + R_s - \mathrm{diag}(R_s))$, which follows from the fact that f_k is an elementwise operation. Note that the second part has diagonal entries equal to zero, since Δ has zero diagonal values.

For the diagonal part $f_k(I_n + \mathrm{diag}(R_s))$, we have

$$\|f_k(I_n + \mathrm{diag}(R_s)) - I_n\|_{\mathrm{op}} = \max_i\{f_k(1 + R_{s,ii}) - 1\} \ .$$

Since $\max_{ij}|R_{s,ij}| = o(1)$ (Equation (28)) and f_k is left continuous at one, we have

$$\max_i\{f_k(1 + R_{s,ii}) - 1\} = o(1) \ .$$

This implies

$$f_k(I_n + \mathrm{diag}(R_s)) = I_n + R_D \ , \quad \|R_D\|_{\mathrm{op}} = o(1) \ .$$

For the off-diagonal part, we define $H = \frac{1}{2}\Delta + R_s - \mathrm{diag}(R_s)$. We apply the Taylor expansion (26) for the off-diagonal entries of $f_k(H)$ to obtain

$$\begin{aligned} f_k(H) &= f_k'(0)H + \frac{1}{2}F \circ H \circ H \\ &= \frac{1}{2}f_k'(0)\Delta + f_k'(0)(R_s - \mathrm{diag}(R_s)) + \frac{1}{2}F \circ H \circ H \,, \\ F_{ij} &= f_k''(\xi_{ij}) \,, \quad |\xi_{ij}| \le |H_{ij}| \,, \quad i \neq j \,, \\ F_{ii} &= 0 \,. \end{aligned}$$

Noticing that $F \circ H \circ H$, the Hadamard product of the matrices, is symmetric, we apply matrix norm inequality (21) to obtain $\|F \circ H \circ H\|_{\mathrm{op}} \le \|F \circ H \circ H\|_\infty$. Moreover,

$$\|F \circ H \circ H\|_\infty = \max_i \sum_{j=1}^n |f_k''(\xi_{ij})| H_{ij}^2 \,.$$

By the smoothness of f_k around zero, we have $\max |f_k''(\xi_{ij})| = O(1)$. Therefore, for any i,

$$\sum_{j=1}^n |f_k''(\xi_{ij})| H_{ij}^2 = O(\sum_{j=1}^n H_{ij}^2) \,.$$

$\sum_j H_{ij}^2$ is the i-th diagonal of H^2, and thus $\sum_j H_{ij}^2 \le \|H\|_{\mathrm{op}}^2$. Based on the derivations above, we obtain $\|F \circ H \circ H\|_{\mathrm{op}} = O(\|H\|_{\mathrm{op}}^2)$. Additionally, based on the definition of H and Equations (27), (28), we have

$$\begin{aligned} \|H\|_{\mathrm{op}}^2 &= O(\|\Delta\|_{\mathrm{op}}^2 + \|R_s\|_{\mathrm{op}}^2 + \|\mathrm{diag}(R_s)\|_{\mathrm{op}}^2) = O(\|\Delta\|_{\mathrm{op}}^2 + \|R_s\|_{\mathrm{op}}^2 + \max_i R_{s,ii}^2) \\ &= O(\|\Delta\|_{\mathrm{op}}^2 + \|R_s\|_{\mathrm{op}}^2) = O(\|\Delta\|_{\mathrm{op}}^2 + \|\Delta\|_{\mathrm{op}}^4) = O(\|\Delta\|_{\mathrm{op}}^2) \,. \end{aligned}$$

To sum up, we show that

$$\begin{aligned} f_k(H) &= \frac{1}{2}f_k'(0)\Delta + R_H \,, \\ R_H &\coloneqq f_k'(0)(R_s - \mathrm{diag}(R_s)) + \frac{1}{2}F \circ H \circ H \,, \quad \|R_H\|_{\mathrm{op}} = O(\|\Delta\|_{\mathrm{op}}^2) \,. \end{aligned}$$

Combining our analysis for diagonal and off-diagonal parts, we obtain our final result

$$\begin{aligned} f_k(\Sigma^{1/2}) &= I_n + \frac{1}{2}f_k'(0)\Delta + R_3 \,, \\ R_3 &\coloneqq R_D + R_H \,, \\ \|R_3\|_{\mathrm{op}} &= o(1) + O(\|\Delta\|_{\mathrm{op}}^2) \,. \end{aligned}$$

C.3 Proof of Proposition 3

Before our proof, we introduce the following matrix norm inequality: For a diagonal matrix A with positive diagonals and $P \succeq 0$, we have

$$\min_i |A_{ii}| \,\mathrm{tr}(P) \le \mathrm{tr}(PA) \le \max_i |A_{ii}| \,\mathrm{tr}(P) \,. \tag{29}$$

We will use this multiple times throughout the proof.

Note that given $Y(k) = X\beta_k$, we can write

$$V(\Sigma) = \frac{K-1}{n} Y(k)^\top f_k(\Sigma) Y(k) = \frac{K-1}{n} \beta_k^\top X^\top f_k(\Sigma) X \beta_k \,.$$

Under the assumption that β_k has zero mean and identity covariance, we have

$$\begin{aligned}\mathbb{E}_{\beta_k} V(\Sigma) &= \frac{K-1}{n}\mathbb{E}\,\mathrm{tr}\big(X^\top f_k(\Sigma) X\beta_k\beta_k^\top\big)\\ &= \frac{K-1}{n}\,\mathrm{tr}\big(X^\top f_k(\Sigma) X\mathbb{E}(\beta_k\beta_k^\top)\big) = \frac{K-1}{n}\,\mathrm{tr}\big(f_k(\Sigma)XX^\top\big)\,.\end{aligned}$$

Note that this holds for any $\Sigma \in \mathcal{E}$.

We write $\Delta = \Sigma - I_n$, which is a symmetric matrix with zero diagonal values. Then we have

$$\begin{aligned}\frac{n}{K-1}(\mathbb{E}_{\beta_k} V(I_n) - \mathbb{E}_{\beta_k} V(\Sigma)) &= \mathrm{tr}\big((f_k(I_n) - f_k(I_n+\Delta))XX^\top\big)\\ &= \mathrm{tr}\big((I_n - f_k(I_n+\Delta))(XX^\top - I_n)\big)\,.\end{aligned} \tag{30}$$

The last equality follows from the fact that $f_k(I_n) - f_k(I_n+\Delta)$ has zero diagonals. Additionally, we have assumed $f_k(1) = 1$ so that $f_k(I_n) = I_n$, as explained in the proof of Theorem 1.

By the elementwise Taylor expansion on f_k (introduced in the proof of Lemma 5), we have

$$f_k(I_n+\Delta) = I_n + f_k'(0)\Delta + \frac{1}{2}R\,, \quad R_{ij} = f_k''(\xi_{ij})\Delta_{ij}^2\,,$$

where ξ_{ij} satisfies $|\xi_{ij}| \le |\Delta_{ij}|$. Then we apply the Taylor expansion to (30) and obtain

$$\begin{aligned}\frac{n}{K-1}(\mathbb{E}_{\beta_k} V(I_n) - \mathbb{E}_{\beta_k} V(\Sigma)) &= \mathrm{tr}\left((-f_k'(0)\Delta - \frac{1}{2}R)(XX^\top - I_n)\right)\\ &= -f_k'(0)\underbrace{\mathrm{tr}\big(\Delta(XX^\top - I_n)\big)}_{=\Delta_f} - \frac{1}{2}\underbrace{\mathrm{tr}\big(R(XX^\top - I_n)\big)}_{=\Delta_R}\,.\end{aligned}$$

Next, we specify Σ to be the solution Σ_η from the one-step PGD-Gauss, and perform matrix analysis to bound Δ_f and Δ_R, respectively. Under the assumption $f_k'(0) \neq 0$, we either have $f_k'(0) > 0$ or $f_k'(0) < 0$. From now on, we assume that $f_k'(0) > 0$ without loss of generality.

Step 1. Analyze Δ_f. By definition of PGD-Gauss, we have

$$\begin{aligned}\Sigma_\eta &= D^{-1}(I_n - \eta f_k'(0)(XX^\top - I_n))^2 D^{-1}\\ &= D^{-1}(I_n - 2\eta f_k'(0)(XX^\top - I_n) + \eta^2(f_k'(0))^2(XX^\top - I_n)^2)D^{-1}\\ &= D^{-2} - 2\eta f_k'(0)D^{-1}(XX^\top - I_n)D^{-1} + \eta^2(f_k'(0))^2 D^{-1}(XX^\top - I_n)^2 D^{-1}\,.\end{aligned}$$

This implies

$$\Delta = \Sigma_\eta - I_n = (D^{-2} - I_n) - 2\eta f_k'(0)D^{-1}(XX^\top - I_n)D^{-1} + \eta^2(f_k'(0))^2 D^{-1}(XX^\top - I_n)^2 D^{-1}\,. \tag{31}$$

Therefore, the difference can be decomposed as

$$\begin{aligned}\Delta_f = &\underbrace{\mathrm{tr}\big((D^{-2} - I_n)(XX^\top - I_n)\big)}_{\text{(I)}} - \underbrace{2\eta f_k'(0)\,\mathrm{tr}\big(D^{-1}(XX^\top - I_n)D^{-1}(XX^\top - I_n)\big)}_{\text{(II)}}\\ &+ \underbrace{\eta^2(f_k'(0))^2\,\mathrm{tr}\big(D^{-1}(XX^\top - I_n)^2 D^{-1}(XX^\top - I_n)\big)}_{\text{(III)}}\,.\end{aligned} \tag{32}$$

It is easy to verify that (I) is zero, since $D^{-2} - I_n$ is a diagonal matrix and $\mathrm{diag}(XX^\top - I_n) = 0$ under Assumption 1. For (II), we apply the inequality (29) with $A = D^{-1}$ and $P = (XX^\top - I_n)D^{-1}(XX^\top - I_n)$

to obtain

$$\begin{aligned}\text{(II)} &\geq \frac{2\eta f_k'(0)}{\max_i D_{ii}} \operatorname{tr}\big((XX^\top - I_n)D^{-1}(XX^\top - I_n)\big) \\ &= \frac{2\eta f_k'(0)}{\max_i D_{ii}} \operatorname{tr}\big(D^{-1}(XX^\top - I_n)^2\big) \ .\end{aligned}$$

Again, we apply (29) with $A = D^{-1}$ and $P = (XX^\top - I_n)^2$ to obtain

$$\text{(II)} \geq \frac{2\eta f_k'(0)}{\max_i D_{ii}^2} \operatorname{tr}\big((XX^\top - I_n)^2\big) = \frac{2\eta f_k'(0)}{\max_i D_{ii}^2} \|XX^\top - I_n\|_F^2 \ . \tag{33}$$

For (III), we define the eigenvalue decomposition $XX^\top = U\Lambda U^\top$ and write

$$\begin{aligned}\operatorname{tr}\big(D^{-1}(XX^\top - I_n)^2 D^{-1}(XX^\top - I_n)\big) &= \operatorname{tr}\big(D^{-1}U(\Lambda - I_n)^2 U^\top D^{-1} U(\Lambda - I_n)U^\top\big) \\ &= \operatorname{tr}\big(U^\top D^{-1} U(\Lambda - I_n)^2 U^\top D^{-1} U(\Lambda - I_n)\big)\end{aligned}$$

We apply inequality (29) with $P = U^\top D^{-1} U(\Lambda - I_n)^2 U^\top D^{-1} U$ and $A = \Lambda - I_n$ to obtain

$$\begin{aligned}\operatorname{tr}\big(D^{-1}(XX^\top - I_n)^2 D^{-1}(XX^\top - I_n)\big) &\leq \max_i |\Lambda_i - 1| \operatorname{tr}\big(U^\top D^{-1} U(\Lambda - I_n)^2 U^\top D^{-1} U\big) \\ &= \max_i |\Lambda_i - 1| \operatorname{tr}\big(D^{-2} U(\Lambda - I_n)^2 U^\top\big) \\ &= \max_i |\Lambda_i - 1| \operatorname{tr}\big(D^{-2}(XX^\top - I_n)^2\big) \ .\end{aligned}$$

Again, we apply (29) with $A = D^{-2}$ and $P = (XX^\top - I_n)^2$ to obtain

$$\begin{aligned}\operatorname{tr}\big(D^{-1}(XX^\top - I_n)^2 D^{-1}(XX^\top - I_n)\big) &\leq \frac{\max_i |\Lambda_i - 1|}{\min_i\{D_{ii}^2\}} \|XX^\top - I_n\|_F^2 \\ &= \frac{\|XX^\top - I_n\|_{\text{op}}}{\min_i\{D_{ii}^2\}} \|XX^\top - I_n\|_F^2 \ .\end{aligned} \tag{34}$$

Since we have analyzed (I), (II), (III), we apply our bounds (33), (34) to (32) and obtain

$$\begin{aligned}\Delta_f &\leq -\frac{2\eta f_k'(0)}{\max_i D_{ii}^2} \|XX^\top - I_n\|_F^2 + \eta^2 (f_k'(0))^2 \frac{\|XX^\top - I_n\|_{\text{op}}}{\min_i\{D_{ii}^2\}} \|XX^\top - I_n\|_F^2 \\ &= \left(-\frac{2\eta f_k'(0)}{\max_i\{D_{ii}^2\}} + \eta^2 (f_k'(0))^2 \frac{\|XX^\top - I_n\|_{\text{op}}}{\min_i\{D_{ii}^2\}}\right) \|XX^\top - I_n\|_F^2 \ .\end{aligned}$$

Step 2. Analyze Δ_R. By definition of R, we have $R_{ii} = 0$ for any i. Hence we have

$$\Delta_R = \sum_{i \neq j} R_{ij} G_{ij} = \sum_{i \neq j} f_k''(\xi_{ij}) \Delta_{ij}^2 G_{ij} \ .$$

where $G_{ij} = X_i^\top X_j$. Under Assumption 1 and the analysis in the proof of Theorem 1, we have $\max |\Delta_{ij}| = o(1)$. In addition, since f_k is smooth around zero, we have $\max |f_k''(\xi_{ij})| = O(1)$. Therefore,

$$\begin{aligned}|\Delta_R| &= O\left(\left|\sum_{i \neq j} \Delta_{ij}^2 G_{ij}\right|\right) \\ &\overset{\text{(i)}}{=} O\left(\sum_{i,j=1}^n \Delta_{ij}^2\right) = O\left(\|\Delta\|_F^2\right) \ .\end{aligned}$$

where (i) follows from $|G_{ij}| \leq 1$ and $\Delta_{ii} = 0$.

Now we give a concrete bound on $\|\Delta\|_F$ for the one-step PGD-Gauss. From (31) and the triangle inequality, we have

$$\begin{aligned}\|\Delta\|_F &= \|(D^{-2}-I_n) - 2\eta f_k'(0) D^{-1}(XX^\top - I_n)D^{-1} + \eta^2 (f_k'(0))^2 D^{-1}(XX^\top - I_n)^2 D^{-1}\|_F \\ &\le \|D^{-2}-I_n\|_F + 2\eta |f_k'(0)| \|D^{-1}(XX^\top - I_n)D^{-1}\|_F + \eta^2 (f_k'(0))^2 \|D^{-1}(XX^\top - I_n)^2 D^{-1}\|_F \ .\end{aligned}$$

Given Equation (23) and Assumption 1, we have $D_{ii}^2 = 1 + O(\eta^2 \|XX^\top - I_n\|_{\mathrm{op}}^2) = 1 + o(1)$, and hence

$$\begin{aligned}\|D^{-2}-I_n\|_F^2 &= \sum_{i=1}^n (1/D_{ii}^2 - 1)^2 = \sum_{i=1}^n \left(\frac{1}{1+\eta^2 \|XX^\top - I_n\|_{\mathrm{op}}^2} - 1 \right)^2 \\ &\le \sum_{i=1}^n \left(\eta^2 \|XX^\top - I_n\|_{\mathrm{op}}^2 \right)^2 = n\eta^4 \|XX^\top - I_n\|_{\mathrm{op}}^4 \ ,\end{aligned}$$

where the last inequality follows from $1 - 1/(1+x) \le x$, for x close to zero. Next, since Equation (23) implies $\max_i |D_{ii} - 1| = o(1)$, we have

$$\|D^{-1}(XX^\top - I_n)D^{-1}\|_F = O(\|XX^\top - I_n\|_F) \ .$$

Lastly, observe that

$$\begin{aligned}\|D^{-1}(XX^\top - I_n)^2 D^{-1}\|_F &= O(\|(XX^\top - I_n)^2\|_F) \\ &= O(\|(XX^\top - I_n)\|_{\mathrm{op}} \|(XX^\top - I_n)\|_F) \ ,\end{aligned}$$

where the last equality follows from $\|AB\|_F \le \|A\|_{\mathrm{op}} \|B\|_F$ for compatible matrices A, B. Combining the results above, we obtain

$$\begin{aligned}\|\Delta\|_F &= O(\sqrt{n}\eta^2 \|XX^\top - I_n\|_{\mathrm{op}}^2) + O(\eta \|XX^\top - I_n\|_F) + O(\eta^2 \|XX^\top - I_n\|_{\mathrm{op}} \|XX^\top - I_n\|_F) \ , \\ |\Delta_R| &= O(\|\Delta\|_F^2) = O(n\eta^4 \|XX^\top - I_n\|_{\mathrm{op}}^4) + O(\eta^2 \|XX^\top - I_n\|_F^2) + O(\eta^4 \|XX^\top - I_n\|_{\mathrm{op}}^2 \|XX^\top - I_n\|_F^2) \ .\end{aligned}$$

Step 3. Derive final results. Based on Step 1 and 2, we have

$$\begin{aligned}\frac{n}{K-1}(\mathbb{E}_{\beta_k} V(I_n) - \mathbb{E}_{\beta_k} V(\Sigma_\eta)) &= -f_k'(0)\Delta_f - \frac{1}{2}\Delta_R \\ &\ge f_k'(0) \left(\frac{2\eta f_k'(0)}{\max_i \{D_{ii}^2\}} - \eta^2 (f_k'(0))^2 \frac{\|XX^\top - I_n\|_{\mathrm{op}}}{\min_i \{D_{ii}^2\}} \right) \|XX^\top - I_n\|_F^2 - \frac{1}{2}\Delta_R \ .\end{aligned} \tag{35}$$

Given Assumption 1 and (23), we have

$$\max_i \{D_{ii}^2\} \asymp 1 \ , \quad \min_i \{D_{ii}^2\} \asymp 1 \ .$$

Thus, we have

$$\begin{aligned}&f_k'(0) \left(\frac{2\eta f_k'(0)}{\max_i \{D_{ii}^2\}} - \eta^2 (f_k'(0))^2 \frac{\|XX^\top - I_n\|_{\mathrm{op}}}{\min_i \{D_{ii}^2\}} \right) \|XX^\top - I_n\|_F^2 \\ &= \Omega\left(f_k'(0) \left(2\eta f_k'(0) - \eta^2 (f_k'(0))^2 \|XX^\top - I_n\|_{\mathrm{op}} \right) \|XX^\top - I_n\|_F^2 \right) \ .\end{aligned}$$

By Assumption 1, it holds that $\eta \|XX^\top - I_n\|_{\mathrm{op}} = o(1)$. Therefore, the term $f_k'(0)\eta \|XX^\top - I_n\|_{\mathrm{op}}$ above is of order $o(1)$. Based on this observation, we further derive

$$f_k'(0) \left(2\eta f_k'(0) - \eta^2 (f_k'(0))^2 \|XX^\top - I_n\|_{\mathrm{op}} \right) \|XX^\top - I_n\|_F^2 = \Omega\left(\eta (f_k'(0))^2 \|XX^\top - I_n\|_F^2 \right) \ .$$

Next we show that Δ_R term is negligible compared to the first term in (35). Based on the analysis in Step 2, we have

$$|\Delta_R| = O(\|\Delta\|_F^2) = O(n\eta^4 \|XX^\top - I_n\|_{\mathrm{op}}^4) + O(\eta^2 \|XX^\top - I_n\|_F^2) + O(\eta^4 \|XX^\top - I_n\|_{\mathrm{op}}^2 \|XX^\top - I_n\|_F^2) \ .$$

and

$$O(\eta^4\|XX^\top - I_n\|_{\mathrm{op}}^2\|XX^\top - I_n\|_F^2) \overset{\text{(i)}}{=} o(\eta^2\|XX^\top - I_n\|_F^2)\ ,$$
$$O(\eta^2\|XX^\top - I_n\|_F^2) \overset{\text{(ii)}}{=} o(\eta\|XX^\top - I_n\|_F^2)\ .$$

In the derivation above, (i) follows from Assumption 1 and (ii) follows from $\eta = o(1)$. Under the additional assumption that $n\eta^3\|XX^\top - I_n\|_{\mathrm{op}}^4 = o(\|XX^\top - I_n\|_F^2)$, we obtain

$$\|\Delta\|_F^2 = o(\eta\|XX^\top - I_n\|_F^2)\ .$$

Therefore,

$$\frac{n}{K-1}(\mathbb{E}_{\beta_k}V(I_n) - \mathbb{E}_{\beta_k}V(\Sigma_\eta)) = \Omega(\eta\|XX^\top - I_n\|_F^2)\ .$$

This completes the proof.

D Quantifying the Gap between Local and Global Optima

As mentioned in Section 4, the solution of the PGD-Gauss is a local optimizer due to non-convex loss. To quantify the gap between local and global optima, we study the competitive ratio defined as below:

$$r := \frac{\|X^\top f(\Sigma^{\mathrm{PGD}})X\|}{\min_{\Sigma\in\mathcal{E}}\|X^\top f(\Sigma)X\|}\ ,$$

where Σ^{PGD} denotes the optimizer obtained from PGD-Gauss. By definition, $r \geq 1$ and smaller r indicates better performance. In words, the competitive ratio quantifies the gap between the local optimal value from PGD-Gauss and the global optimal value, and such type of analysis, generally known as competitive analysis, is widely used in the online learning literature [Borodin and El-Yaniv, 2005] to quantify the optimality gap between an online algorithm and an optimal baseline. Our competitive ratio follows a similar idea, but is applied to a non-convex offline optimization.

For design optimization, achieving $r \approx 1$ is hard, as the global optimization relates to NP-hard problems (Section 1.1). Alternatively, we lower bound the global optimum using a semidefinite relaxation:

$$\min_{\Sigma\in\mathcal{E}}\|X^\top f(\Sigma)X\| = \min_{C=f(\Sigma),\Sigma\in\mathcal{E}}\|X^\top CX\| \geq \min_{C\succeq 0, C_{ii}=f(1)}\|X^\top CX\|.$$

The right-hand side is convex and thus can be efficiently solved, leading to an upper bound of r:

$$r \leq \frac{\|X^\top f(\Sigma^{\mathrm{PGD}})X\|}{\min_{C\succeq 0, C_{ii}=f(1)}\|X^\top CX\|}.$$

Therefore, one can always compute a numerical bound on the competitive ratio to assess the gap between the local and global optima. It remains an open question to us if a theoretical bound on the competitive ratio can be derived without stringent conditions on the loss landscape. We consider it as future work.

E Covariate-adaptive Designs and Covariate Adjustments

Our work has focused on optimizing the MSE property of Horvitz-Thompson estimators. In practice, researchers in the analysis stage could utilize more advanced estimators with covariate adjustments. For instance, Lin's estimator [Lin, 2013] is widely used under the binary treatment setting, which is defined as the coefficients of D_i in the OLS regression

$$Y_i \sim X_i + D_i + D_i(X_i - \bar{X})\ , \quad \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i\ .$$

Li and Ding [2017] show the "optimality" of Lin's estimator within a class of regression-adjusted estimators. Here we clarify two practical questions related to covariate adjustments and Gaussianization:

(1) Why do researchers consider covariate-adaptive designs? Alternatively, one could simply use covariate adjustments under complete randomization.
(2) Suppose we want to balance for covariates in the design stage. What if we formulate covariate balance measures with respect to covariate-adjusted estimators?

First, in real-world randomized experiments, designers and analyzers may not communicate or share the same set of covariates. Therefore, balancing for covariates in the design stage is desirable, as it improves the estimation precision even with the simple Horvitz-Thompson estimator. Moreover, studies have shown that covariate-adaptive designs —such as rerandomiztion— never hurt the estimation [Li and Ding, 2020]. We anticipate similar results hold under Gaussianization. That is, we view covariate-adaptive designs and covariate adjustments as synergistic approaches that can be combined to further enhance the estimation.

Second, under Gaussianization, it is possible to analyze the MSE property of covariate-adjusted estimators as in Chang [2023] and formulate covariate balance measures tailored to these estimators. However, we argue that design optimization is more robust toward different outcome-generating models when using model-agnostic estimators, e.g., Horvitz-Thompson estimators. In other words, under a misspecified model, the MSE of covariate-adjusted estimator will also be misspecified, and the estimator itself might be biased. Hence, conducting design optimization based on biased adjustments could impair estimation precision.

F Discussions on the Gaussianized Design Class

As mentioned in Section 1.1, Gaussianization imposes constrained design class, which is, in general, a subset of all possible designs. Here, we provide more explanations on the Gaussianized design class by listing the designs that can be Gaussianized, and those that cannot. For simplicity, we focus on the binary treatment setting, that is, $D_i \in \{1, 2\}$.

- Designs that can be Gaussianized: i.i.d. Bernoulli design, matched pair design, and certain instances of rerandomization (RR) and Gram-Schmidt-Walk (GSW) design.
- Designs that cannot be Gaussianized: complete randomization, and other instances of RR and GSW design.

Concretely, Proposition 4 formally shows that the matched pair design can be Gaussianized.

Proposition 4. *Suppose H is a block diagonal matrix with each block equal to*

$$\frac{1}{4}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},$$

That is, $H = \mathrm{Cov}_1(D) = \mathrm{Cov}_2(D)$ for D generated from a certain matched pair design. Then, we have

$$H = \mathrm{Cov}_1(D) = f_1(4H)\,, \quad H = \mathrm{Cov}_2(D) = f_2(4H)\,, \quad 4H \in \mathcal{E}\,.$$

That is, $\mathrm{Cov}_k(D)$ can be Gaussianized by $D = g(T), T \sim \mathcal{N}(0, 4H)$.

Proof. By specifying Proposition 1 to the $K = 2$ case, we have

$$f_1(\rho) = f_2(\rho) = r_{1,1}(\rho) = \int_0^\rho \frac{1}{2\pi\sqrt{1-r^2}}\mathrm{d}r = \frac{1}{2\pi}\arcsin(\rho)\,.$$

Based on the analytical expression above, we have

$$\frac{1}{4} = f_1(1)\,, \quad -\frac{1}{4} = f_1(-1)\,, \quad 0 = f_1(0)\,.$$

Therefore, we have $H = f_1(4H)$, which completes the proof. □

Proposition 4 leverages the specific form of f_k in the binary treatment setting, and the argument for other designs can be proved in a similar manner. Qualitatively, a design can be Gaussianized when its covariance matrix is either full-rank (e.g., i.i.d. Bernoulli design) or relatively low-rank (e.g., the matched pairs design). Consequently, some instances of RR and GSW designs can be Gaussianized, while some cannot. When the design is Gaussianizable, one can apply our method to improve the design by initializing the PGD-Gauss from its Gaussianized representation. As we explained in Section 1.1, although PGD only finds local optimum, one can flexibly initialize.

When a design cannot be Gaussianized, one may compute an approximate Gaussianization by solving

$$\min_{\Sigma \in \mathcal{E}} \|C - f(\Sigma)\|_{\text{norm}} .$$

Here C is the covariance matrix corresponding to the given design, and f is the Gaussianization mapping obtained from Proposition 1. For instance, suppose the rerandomization induces a design D that cannot be Gaussianized. Then, we may solve for

$$\min_{\Sigma \in \mathcal{E}} \|\text{Cov}_1(D) - f_1(\Sigma)\|_{\text{norm}} ,$$

with the function f_1 obtained in Proposition 4. That is, one can initialize PGD-Gauss from an approximate Gaussianization, which further mitigates the limitations of Gaussianization.