cemmap
centre for microdata methods and practice

Robust signal maximization in spillover experiments

Kirill Borusyak
Peter Hull
Evan Munro

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP13/26

Economic and Social Research Council

Robust Signal Maximization in Spillover Experiments

Kirill Borusyak
UC Berkeley

Peter Hull
Brown

Evan Munro
Chicago Booth*

July 2026

Abstract

We study the optimal design and analysis of experiments for estimating spillover effects. Assuming a known (e.g., linear) exposure mapping, we characterize the treatment-assignment distribution and regression-based estimator that minimize worst-case asymptotic variance against a broad class of distributions of unobservables. The design problem yields an intuitive solution in which the planner trades off spillover signal strength against diffusion of spillover variation. The analysis problem yields a simple recentered instrumental variable estimator to best leverage this variation. This framework produces natural solutions in several benchmark cases—such as clustered exposure—and suggests computationally tractable approximations for general networks, including bipartite settings. We illustrate these new tools in semi-synthetic experiments based on two applications from development economics. Our approach yields large standard error reductions in both experiments, increasing effective sample sizes by 50–100% or more.

*Contact: k.borusyak@berkeley.edu, peter_hull@brown.edu, and evan.munro@chicagobooth.edu. We thank David Atkin, Michael Best, and Eric Verhoogen for encouraging this project; we thank Gabriel Kreindler, Shuangning Li, and Davide Viviano for useful comments. OpenAI's ChatGPT contributed valuable insights. Refine.ink was used to check for consistency and clarity.

1 Introduction

Researchers increasingly estimate spillovers in experiments. Sometimes, an intervention is randomized to one set of units while the spillover effects are measured in another set. In such “bipartite” experiments the intervention and outcome units are connected via a network: e.g., of friendship across individuals (Cai et al., 2015) or of supplying relationships across firms (Best et al., 2025). In other cases, the intervention and outcome units coincide and direct and spillover effects are jointly estimated in a single experimental sample (Miguel and Kremer, 2004; Chaurey et al., 2025). Power is a common issue in such studies: spillover effect estimates can be noisy, while increasing the sample size is costly. It is therefore important to make the best use of the experimental variation, as well as to design the randomization to target the spillover parameter of interest.

This paper develops new tools for experimental design and estimation when spillover effects are of primary interest. Our baseline assumption is that the researcher adopts a linear model of spillovers, in which the key treatment variable is the fraction or number of “friends” (i.e., network connections) who have been randomly selected for the intervention—perhaps augmented with weights reflecting the importance of each connection.[1] Such linear specifications are widespread in economics, including in experiments (e.g., Miguel and Kremer, 2004; Cai et al., 2015; Atkin et al., 2024). We assume the researcher has access to the network measure (or at least a correlated proxy) at the design stage. We further assume the researcher wants to identify the spillover effect using the experimental variation; hence, they use some recentered estimator (Borusyak and Hull, 2023). It is increasingly recognized that recentering is necessary to guarantee a causal interpretation of spillover effect estimates without assuming that the network is exogenous, with appropriate regression or instrumental variable (IV) estimators employed either intuitively (Miguel and Kremer, 2004) or formally (Chaurey et al., 2025; Atkin et al., 2024).

We then ask: which randomization design and which recentered estimator are likely to yield good power? Our answer reflects two countervailing intuitions. On the one hand, when an outcome unit i is exposed to the shocks of multiple intervention units k, it is desirable to positively correlate the treatment assignments of those intervention units. Doing so increases the variance of i’s spillover treatment (i.e., the spillover “signal”). On the other hand, when multiple outcome units are exposed to the same intervention unit, their exposures are correlated and this can increase estimation noise if their unobservables are also correlated; correlating assignments of multiple intervention units exacerbates this clustering problem. This makes it more desirable to increase the degree of independence across different exposures, spreading experimental variation out across the sample (“isotropy”). A likely-powerful design is thus one which maximizes spillover signals while ensuring spillover variation is sufficiently diffuse. The estimator choice can also help with isotropy, reweighting the data to spread out the induced spillover variation, at the cost of a lower signal.

We formalize these intuitions in a minimax approach: we look for the experimental design and

[1] Specifically, we assume the spillover treatment can be written $x_i = \sum_k w_{ik} g_k$ where g_k is an indicator of unit k assigned to the treatment group. The w_{ik} are fully unrestricted, and the intervention units $k = 1, \ldots, K$ can differ from the outcome units $i = 1, \ldots, N$. We also discuss how our approach can be used for nonlinear formula treatments $x_i = X_i(g_1, \ldots, g_K)$ for known $X_i(\cdot)$ that can implicitly depend on some w_{ik} (Borusyak et al., 2025b).

recentered estimator that minimize the worst-case approximate variance over a large class of error distributions. We parameterize this class parsimoniously, by a scalar that captures the extent to which the researcher is willing to rule out mutual dependencies, dependence on the network, and heteroskedasticity of the errors. This approach has three advantages. First, it reflects the reality that researchers tend to have at best a rough sense of how errors are clustered, especially at the experimental design stage. Second, it allows for general network structures and yields a tractable closed-form characterization of worst-case variance, both as a function of the estimator (holding the design fixed) and as a function of the design (provided the optimal estimator will be used). Although implementing our solution generally involves numerical optimization, it is amenable to practical algorithms. Finally, the minimax approach yields non-degenerate and intuitive solutions in contrast to some corner solutions in the literature (e.g., Kiefer (1974)). For example, independent randomization and cluster-level randomization arise as special cases when, respectively, there are no spillovers and when the exposure is the cluster-level treatment saturation. When exposure is a leave-one-out average of cluster shocks, the solution is to optimally mix between these two designs.

We characterize the minimax-optimal design and estimator in two steps. First, we derive the optimal recentered IV estimator for any experimental design. Its construction entails taking the spillover treatment itself, recentering it by its expectation over the experimental design, and partially "whitening" it to spread out the remaining exogenous variation. While the whitening step weakens the first stage, the induced isotropy improves robustness against non-spherical errors. Second, we characterize the best designs given the optimal recentered instrument will be used. We show this optimal design satisfies natural properties, such as a 50/50 marginal distribution for each shock and separability across independent network blocks. The optimal IV is straightforward to construct for any design, and we propose a computationally efficient approximation to the optimal design; the entire procedure is collected in Algorithm 1, below. In the parameterization with the highest robustness to adversarial errors, this procedure yields an especially intuitive closed-form solution in which the correlation of shock assignments is determined by the cosine similarity of exposure to them. We show how our estimators are asymptotically normal with a sparse exposure network—even when the errors have strong mutual correlations—and give a simple inference procedure.

We extend these solutions in several directions, including by characterizing minimax-optimal designs and estimators when spillover effects are estimated alongside direct effects or other linear exposure measures. Here the optimal design balances signal and isotropy of the residual variation in the focal spillover treatment after accounting for the other exposures. Other extensions include estimators with predetermined covariates, nonlinear spillover treatments, and optimal design when a researcher faces budget constraints in allocating shocks.

We then illustrate the power gains from using our approach, relative to independent randomization, in semi-synthetic experiments calibrated to the data from the bipartite experiment in Cai et al. (2015) and the joint estimation of direct and spillover effects in Miguel and Kremer (2004). We find large declines in standard errors in both settings, equivalent to increasing effective sample sizes by around 50–100% or more relative to the independent randomization in the original papers and no whitening. These gains arise from both the optimized experimental design and the use of optimal recentered instruments.

This paper contributes to several related literatures. Most directly, we add to a growing literature on optimal experimental design under interference. Many papers in this literature focus on specific settings, a restricted class of designs, and a pre-specified estimator (with different models of interference and estimands). For example, Baird et al. (2018) and Cruces et al. (2025) focus on partial interference, where spillovers are confined within clusters, and derive optimal saturation designs for estimating direct and spillover effects, including by regression. Pouget-Abadie et al. (2019) and Harshaw et al. (2023) instead focus on bipartite experiments and derive optimal clustered assignments under a linear exposure-response model with heterogeneous treatment effects. Thiyageswaran et al. (2026) study optimal worst-case estimation of the global average treatment effect (GATE) over a general class of designs using the standard Horvitz-Thompson estimator, which does not account for spillovers and is generally biased. Faridani and Niehaus (2026) study experimental design and estimation of the GATE when the structure of interference is unknown but decays with distance sufficiently fast, deriving minimax convergence rates.[2]

Relative to this literature, our approach differs in four ways. First, we provide a unified and tractable characterization of optimal experimental designs for general network structures, nesting partial interference and bipartite graphs as special cases. Second, we do not restrict the class of designs *a priori*. Third, we jointly optimize the design and the estimator to achieve further power improvement. We minimize the finite-sample approximate variance of the estimator rather than only its convergence rate, as in Faridani and Niehaus (2026). Like Harshaw et al. (2023), we employ recentering to get unbiased estimators without assuming network exogeneity. Finally, we target the parameters of a linear spillover model, rather than GATE-style estimands. While more restrictive, such constant-effects models are widely used in practice with evidence for linearity in different contexts (e.g., in Miguel and Kremer (2004) and Egger et al. (2022)).[3] When both direct and indirect effects are included, our approach focuses on isolating them rather than obtaining the total in a GATE-style analysis. Our problem can also be reformulated to target the GATE under a known linear-exposure structure. This additional structure is expected to deliver more power than designs targeting GATE under more non-parametric (linear or nonlinear) interference, e.g. in Faridani and Niehaus (2026).

We also add to a literature using minimax arguments to justify different experimental designs. In settings without interference, Kallus (2021) and Bai (2023) show that complete or blockwise randomization is minimax-optimal absent strong assumptions on error dependence. We extend this logic to spillover estimation with recentered IV, where we show how the minimax-optimal design depends on the allowed adversarial clustering of errors. As in Thiyageswaran et al. (2026), we impose a Schatten p-norm constraint on the second-moment matrix of model residuals to define

[2] Other work with general networks includes Viviano (2026) who studies two-wave experiments for estimands including average treatment and spillover effects, using a pilot wave to estimate error variances, and Viviano et al. (2026) who choose treatment clusterings to trade off worst-case asymptotic bias and variance when estimating GATEs.

[3] In the presence of heterogeneous treatment effects, our baseline estimators identify their convex averages. This is shown in Appendix B.1, in line with earlier results on design-based estimators (Borusyak and Hull, 2024).

our minimax problem; choosing a smaller p leads to greater robustness to cross-unit dependence and pushes the optimal design towards more diffuse randomization. The ex-ante choice of p by the researcher is similar to choices made in the broader literature on minimax-optimal estimators.[4]

The minimax approach allows us to avoid degenerate solutions that often arise in other analyses. For example, a classic literature on optimal regression design (e.g., Kiefer, 1974) chooses where to place observations in covariate space to minimize a functional of the ordinary least squares (OLS) covariance matrix (e.g. D-optimality, which minimizes its determinant; Kiefer and Wolfowitz (1959)). This literature assumes independent and homoskedastic errors, and optimal designs tend to be concentrated on a few extreme points in the design space.[5] In the interference setting, the optimal design in Pouget-Abadie et al. (2019) can similarly produce very coarse randomization: e.g., grouping all units into two clusters only and assigning treatment at the cluster level. Harshaw et al. (2023) employ a heuristic correction to the optimization criterion to reduce the issue, while our minimax approach resolves it naturally—yielding a guarantee of asymptotic normality with a sparse exposure network and mild regularity conditions.

Finally, we add to a literature on robust and powerful spillover estimation with recentered IV. Borusyak and Hull (2023) show how recentering via knowledge of the (quasi-)experimental design can identify spillover effects under arbitrary endogeneity of the network, while Borusyak and Hull (2026) and Borusyak et al. (2025a) show how asymptotically optimal recentered instruments can be constructed given a particular design. Here we optimize both the design and instrument, with our applications showing that both levers can meaningfully improve estimator precision. Moreover, while the implementation of the Borusyak and Hull (2026) optimal estimator is hindered by its dependence on the (likely unknown) dependence of error terms, our minimax solution provides a parsimonious solution robust to a wide range of error structures.

The remainder of this paper is organized as follows. In Section 2 we develop our theoretical framework and results. In Section 3, we propose tractable methods of computing the optimal design and provide a central limit theorem for the recentered IV estimator under the feasible design. In Section 4 we illustrate these tools in two semi-synthetic experiments. Section 5 concludes. The proofs of main results are given in Appendix A. Additional results are given in Appendix B, with proofs in Appendix C.

2 Theory

2.1 Setting

We consider estimation of the spillover effects of a set of treatment shocks $g_k \in \{0,1\}$, $k = 1, \ldots, K$, on a set of outcomes y_i, $i = 1, \ldots, N$. This might be in a bipartite setting, in which the K intervention units are distinct from the N outcome units, or in a setting like Miguel and Kremer

[4] In non-parametric regression, for example, researchers choose the smoothness class of the data-generating process—determining the tradeoff between robustness and confidence interval length (Armstrong and Kolesár, 2018).

[5] Our focus on non-spherical errors echoes Bickel and Herzberg (1979) who recover non-degenerate designs when errors are serially correlated.

(2004) in which we estimate spillovers within a single experimental sample (with $i = k$).

We parameterize spillovers with a linear exposure model, which we assume is correctly specified:

$$y_i = \beta x_i + \varepsilon_i, \qquad x_i = w_i' g, \tag{1}$$

where the spillover treatment x_i combines the shocks $g = (g_1, \ldots, g_K)'$ with exposure weights $w_i \in \mathbb{R}^K$, which may or may not add to one, and ε_i is an unobserved error.[6] For example, Cai et al. (2015) estimate a spillover effect β from information sessions randomized to a set of K farmers in China on their peers' decisions to adopt weather insurance y_i; here $x_i = w_i' g$ gives the share of farmer i's neighbors who were assigned to the information session, with w_{ik} being entries of the row-normalized adjacency matrix of the peer graph.[7] In Section 2.4 we extend the model to include direct effects of treatment, for settings like Miguel and Kremer (2004). Below we discuss how our main results extend to spillover treatments that are nonlinear functions of the shocks.

In most of our analysis, we condition on the exposure matrix $W = (w_1', \ldots, w_N')'$; we make this conditioning implicit to simplify notation. We allow the error vector $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_N)'$ to be either fixed (as in a standard design-based setup) or drawn from some distribution that can implicitly depend on W.[8] Without loss of generality, we assume that no column of W is entirely zero.

We consider the problem of choosing an experimental design, i.e., a probability distribution for g: $\delta \in \mathcal{D}$ where $\mathcal{D} = \Delta(\{0,1\}^K)$ and $\Delta(\cdot)$ denotes the simplex. By virtue of randomization, we have $g \perp\!\!\!\perp \varepsilon$. We then estimate β by recentered IV (Borusyak and Hull, 2023): i.e., we choose a set of instrument functions $z_i(\cdot) : \{0,1\}^K \to \mathbb{R}$ where $\mathbb{E}_\delta [z_i(g)] = 0$ for all i, and estimate:

$$\hat{\beta}[z] = \frac{\sum_{i=1}^N z_i(g) y_i}{\sum_{i=1}^N z_i(g) x_i}.$$

We let $\mathcal{Z}_\delta$ be the set of all such $z = (z_i)_{i=1}^N$ for a given δ.

Recentering (i.e., the mean-zero property of $z_i(g)$, implicitly conditional on W) ensures we identify β just using the experimental variation in g: the network W can be arbitrarily correlated with the unobserved ε. Indeed, the moment condition $\mathbb{E}_\delta \left[\sum_{i=1}^N z_i \varepsilon_i\right] = 0$ is satisfied for any such z. Recentered $z \in \mathcal{Z}_\delta$ can be obtained from any fixed function, $\tilde{z}_i$, of both g and W by subtracting its (known) expectation under a given design: i.e. $z_i = \tilde{z}_i - \mathbb{E}_\delta [\tilde{z}_i]$. The set of recentered IV estimators

[6] Here correct specification includes an implicit exclusion restriction: that the shocks only affect outcomes through x_i. While we focus on a linear model with constant effects, Appendix B.1 discusses the interpretation of our proposed estimands under heterogeneous treatment effects.

[7] We assume that the relevant network is measured at the design stage, but require only that we observe the connections between outcome and intervention units. We do not need, for example, connections between one intervention unit and another at the design stage. This is not a strong requirement, as the connections between outcome and intervention units are necessary to estimate the regression in (1) anyway (although not necessarily at the design stage). Efficiency gains are also likely if the design is based on a noisy measure of the network, measured at baseline. If the network can itself endogenously change after the experiment, the researcher can use the ex-post network for the endogenous variable in the regression while using the ex-ante network to construct the IV, as in Gao (2024).

[8] All of our results hold in both cases but the restriction on the error distribution introduced below is more natural when ε are viewed as stochastic. Nevertheless, our approach is still design-based in the sense that the moment conditions we leverage derive their validity from the randomization of g alone.

$\hat{\beta}[z]$, $z \in \mathcal{Z}_\delta$, is large, including estimators that weight by functions of W (by rescaling z), control for functions of W (by the Frisch–Waugh–Lovell theorem), as well as non-instrumented estimators (e.g., OLS of y_i on $\tilde{z}_i = x_i$ controlling for $\mathbb{E}_\delta[\tilde{z}_i]$, again by the Frisch–Waugh–Lovell theorem).[9]

With this setup, we ask: what design δ and instrument z should we choose in order to get the most precise estimate of the spillover effect β?

2.2 Optimal Design and Instrument

We follow Borusyak and Hull (2026) in studying the finite-sample approximate variance of the recentered IV estimator:

$$\mathcal{V}_{\delta,\mathcal{E}}[z] = \frac{\mathrm{Var}_{\delta,\mathcal{E}}\left[\sum_{i=1}^N z_i(g)\varepsilon_i\right]}{\mathbb{E}_\delta\left[\sum_{i=1}^N z_i(g)x_i\right]^2},$$

where here we let $\mathcal{E}$ denote the (unknown) distribution of ε, with $\mathcal{V}_{\delta,\mathcal{E}}[z] = \infty$ whenever the denominator is zero. Proposition 2 of Borusyak and Hull (2026) shows that $\mathcal{V}_{\delta,\mathcal{E}}[z]$ gives a good approximation to the appropriately scaled asymptotic variance of $\hat{\beta}[z]$ so long as the estimator is well-behaved in a particular sense, which primarily means it converges to β at some rate with a non-vanishing first stage.[10]

We look for the experimental design and instrument that minimize the worst-case $\mathcal{V}_{\delta,\mathcal{E}}[z]$ under a weak restriction on the error distribution. Specifically, we suppose the researcher believes $\mathcal{E}$ belongs to some class of distributions $\mathcal{F}_p(\sigma)$ and solve the minimax problem:

$$\delta^*, z^* \in \arg\min_{\delta\in\mathcal{D}, z\in\mathcal{Z}_\delta} \max_{\mathcal{E}\in\mathcal{F}_p(\sigma)} \mathcal{V}_{\delta,\mathcal{E}}[z]. \tag{2}$$

The class of error distributions, parameterized by $p \in [1, \infty]$ and $\sigma > 0$, is given by:

$$\mathcal{F}_p(\sigma) = \left\{\mathcal{E}\colon \left\|\mathbb{E}_\mathcal{E}\left[\varepsilon\varepsilon'\right]\right\|_p \leq N^{1/p}\sigma^2\right\}, \tag{3}$$

where $\|\cdot\|_p$ is the Schatten-p matrix norm defined for positive semi-definite matrices as[11]

$$\|A\|_p = \left(\mathrm{tr}\left(A^p\right)\right)^{1/p}.$$

Three important special cases are the nuclear norm (a.k.a., the trace norm) with $p = 1$, the Frobenius

[9] Appendix B.2 considers estimation without recentering. While non-recentered IVs can in principle improve on mean-squared error grounds by introducing a small amount of bias to estimation, we show the corresponding optimal designs can have undesirable properties like unusually high treatment rates. We also show that when $W\mathbf{1} = 1$ and an intercept is included in estimation (or, more generally, when $W\mathbf{1}$ is linearly spanned by the included controls), there is no loss in only considering recentered instruments since any bias in non-recentered instruments is absorbed.

[10] As a just-identified IV estimator, $\hat{\beta}[z]$ may have no finite-sample moments. More precisely, in our setting with discrete g the problems may arise from $\hat{\beta}[z]$ not well-defined in a small-probability event that $\sum_{i=1}^N z_i(g)x_i = 0$.

[11] For general matrices, it is defined as $\left(\mathrm{tr}\left(\sqrt{A'A}\right)^p\right)^{1/p}$. Schatten norms should not be confused with other matrix norms that can be denoted in the same way.

norm with $p = 2$, and the operator norm with $p = \infty$. Denoting the eigenvalues of $\mathbb{E}_{\mathcal{E}}[\varepsilon\varepsilon']$ by $\lambda_1, \ldots, \lambda_N \geq 0$, we can rewrite (3) as an upper bound on their power mean:

$$\mathcal{F}_p(\sigma) = \left\{ \mathcal{E} \colon \left(\frac{1}{N} \sum_{i=1}^{N} \lambda_i^p \right)^{1/p} \leq \sigma^2 \right\},$$

with the limit $\mathcal{F}_p(\sigma) = \left\{ \mathcal{E} \colon \lambda_{\max}\left(\mathbb{E}_\delta[\varepsilon\varepsilon']\right) \leq \sigma^2 \right\}$ where $\lambda_{\max} = \max\{\lambda_1, \ldots, \lambda_N\}$ for $p = \infty$.

To interpret $\mathcal{F}_p(\sigma)$, note first that spherical (i.e., homoskedastic and mutually uncorrelated) errors with mean zero (implicitly conditional on W) and variance σ^2, i.e. $\mathbb{E}_{\mathcal{E}}[\varepsilon\varepsilon'] = \sigma^2 I_N$, are at the boundary of this set for any p. Different choices of p then place different weights on the correlations among the errors. When $p = 1$, only the average second moment of ε_i is constrained, allowing fully unrestricted correlations across observations. At the other extreme, when $p = \infty$, correlations are strongly penalized such that a worst-case scenario in (2) is always with spherical errors. Intermediate values of p correspond to smaller penalties on correlations. Figure 1 provides a visualization for $N = 2$ by plotting the boundary of $\mathcal{F}_p(\sigma)$ in terms of the two eigenvalues of $\mathbb{E}_{\mathcal{E}}[\varepsilon\varepsilon']$; increasing the correlation across errors makes the eigenvalues more asymmetric without changing their sum; see also Appendix B.3 for analytical results with equicorrelated ε_i. Overall, p captures the extent to which the researcher is willing to rule out strong correlations across the errors when designing the experiment. Similarly, deviations from homoskedasticity and from network exogeneity (i.e., the zero conditional mean of the errors) are penalized more strongly when p is larger.

Optimal Instrument. Solving the problem (2) backwards, we first characterize the optimal instrument for a given design δ:

Theorem 1. *Fix δ and let*

$$\tilde{x}_\delta = W\tilde{g}_\delta, \qquad \tilde{g}_\delta = g - \mathbb{E}_\delta[g],$$
$$S_\delta = \mathrm{Var}_\delta[x] = W \mathrm{Var}_\delta[g] W'.$$

Suppose $S_\delta \neq 0$. Then for any $p \in [1, \infty]$ and $\sigma > 0$, the optimal instrument is:

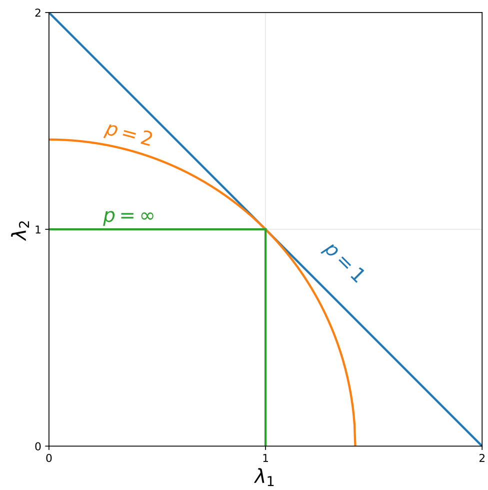
$$z_\delta^*(g) \propto \left(S_\delta^\dagger\right)^{\frac{1}{p+1}} \tilde{x}_\delta \in \arg\min_{z \in \mathcal{Z}_\delta} \max_{\mathcal{E} \in \mathcal{F}_p(\sigma)} \mathcal{V}_{\delta,\mathcal{E}}[z],$$

where $\dagger$ denotes the pseudoinverse. The worst-case approximate variance is:

$$\max_{\mathcal{E} \in \mathcal{F}_p(\sigma)} \mathcal{V}_{\delta,\mathcal{E}}[z_\delta^*] = N^{1/p}\sigma^2 \left(\mathrm{tr}(S_\delta^{p/(p+1)})\right)^{-(p+1)/p}.$$

The Appendix A.1 proof provides an explicit characterization of the worst-case error distribution. In this proposition and for the rest of the paper, we use the convention that $1/\infty = 0$, $p/(p+1) = 1$ for $p = \infty$, and $A^0 = A^\dagger A$ for a matrix A.

Figure 1: Boundary of $\mathcal{F}_p(\sigma)$ for $\sigma^2 = 1$, $N = 2$, for Different Choices of p



Notes: This figure shows boundaries of $\mathcal{F}_p(\sigma)$ for different Schatten-p norms, with $N = 2$ observations and $\sigma = 1$, in terms of the two eigenvalues of the $\mathbb{E}_{\mathcal{E}}[\varepsilon\varepsilon']$ matrix. The $(1, 1)$ point corresponds to spherical errors.

Intuitively, z^*_δ takes the spillover treatment x, recenters it by $\mathbb{E}_\delta[x]$, and then reweights the resulting $\tilde{x}_\delta$ inversely by a fractional power of $\text{Var}_\delta[x]$, $S_\delta^{1/(p+1)}$. The goal in reweighting is to spread out the useful variation in $\tilde{x}_\delta$ across different directions that could be adversarially targeted by clustering in ε. When $p = 1$ and $\text{Var}_\delta[x]$ is an invertible matrix, this fully "whitens" $\tilde{x}_\delta$ (i.e., rotates it to make $\text{Var}_\delta[z^*_\delta] = I$) in order to guard against the fully-unconstrained clustering of ε,[12] while as p increases we do less whitening as errors with strong mutual correlations are excluded. As a result, at $p = \infty$, the optimal IV is the unwhitened $\tilde{x}_\delta$.[13]

Optimal Design. We next characterize the optimal design, assuming the optimal instrument will be used for estimation:

Proposition 1. *Define $\tilde{x}_\delta$ and S_δ as in Theorem 1 and suppose there is $\delta \in \mathcal{D}$ such that $S_\delta \neq 0$. Then*

$$\delta^*, z^* \in \arg\min_{\delta\in\mathcal{D}, z\in\mathcal{Z}_\delta} \max_{\mathcal{E}\in\mathcal{F}_p(\sigma)} \mathcal{V}_{\delta,\mathcal{E}}[z].$$

is solved by

$$\delta^* \in \arg\max_{\delta\in\mathcal{D}} \text{tr}\left(S_\delta^{\frac{p}{p+1}}\right)$$

together with the optimal instrument $z^* = z^*_{\delta^*}$ from Theorem 1.

This result follows directly from Theorem 1. For $p = 1$, the optimal design maximizes $\text{tr}\left(\text{Var}_\delta[x]^{1/2}\right)$. At $p = \infty$, the sum of variances, $\text{tr}(\text{Var}_\delta[x])$, is maximized instead.

[12] When $\text{Var}_\delta[x]$ is not invertible, e.g. because δ perfectly correlates x_i for several i, z^* preserves that dependence.

[13] Theorem 1 technically only shows this for invertible $\text{Var}_\delta[x]$; the general optimality of $\tilde{x}_\delta$ for $p = \infty$, i.e. when $\mathcal{F}_p(\sigma)$ bounds the operator norm of the error second-moment matrix, was previously shown in Borusyak and Hull (2026, Lemma 2).

In what follows we avoid $p = \infty$ as the optimal design problem has a degenerate rank-1 solution. Specifically, Appendix B.4 shows that there exists δ^* that puts all the mass on just one pair of complementary assignments $g^* \in \{0,1\}^K$ and $\mathbf{1} - g^*$. Moreover, if W is non-negative, one solution corresponds to $g^* = \mathbf{1}$, i.e. treating all intervention units with probability 0.5 and treating none otherwise. While this is not problematic under spherical errors, it is highly undesirable when errors can be correlated since the estimator may not be consistent. Choosing $p < \infty$ guards against this possibility and helps with consistency, as we show in Section 3.3.[14]

A further result helps characterize optimal designs:

Proposition 2. *There is a δ^* from Proposition 1 with $\mathbb{E}_{\delta^*}[g_k] = 0.5$ for all k. Further, if W is block-diagonal after a joint partition of rows and columns, there is such a δ^* with independent subvectors of g across blocks and the optimal design can be solved separately block-by-block.*

Intuitively, any design with $\mathbb{E}_{\delta^*}[g_k] \neq 0.5$ can be symmetrized by averaging it with its complement and this can never hurt the criterion. Similarly, any across-block dependence in g_k can get removed without hurting the criterion, since $S_{\delta^*} = W \mathrm{Var}_{\delta^*}[g] W'$.[15] This result is generalized in Appendix B.5: whenever W is invariant under a group of relabelings over rows and columns, there exists a δ^* which is invariant under the same group. In particular, if W is invariant to permutations of rows and columns in a cluster, one may restrict attention to exchangeable designs within that cluster.

2.3 Special Cases

We now illustrate the results of Theorem 1 and Proposition 1 in several insightful special cases.

No Spillovers. A useful benchmark is the standard experimental case with $N = K$ and $W = I$. Since W is diagonal, Proposition 2 immediately implies that independent random assignment, $g_i \overset{iid}{\sim}$ *Bernoulli*(0.5), is optimal in this case, and the optimal instrument is just $z_i^*(g) \propto g_i - 0.5$.[16]

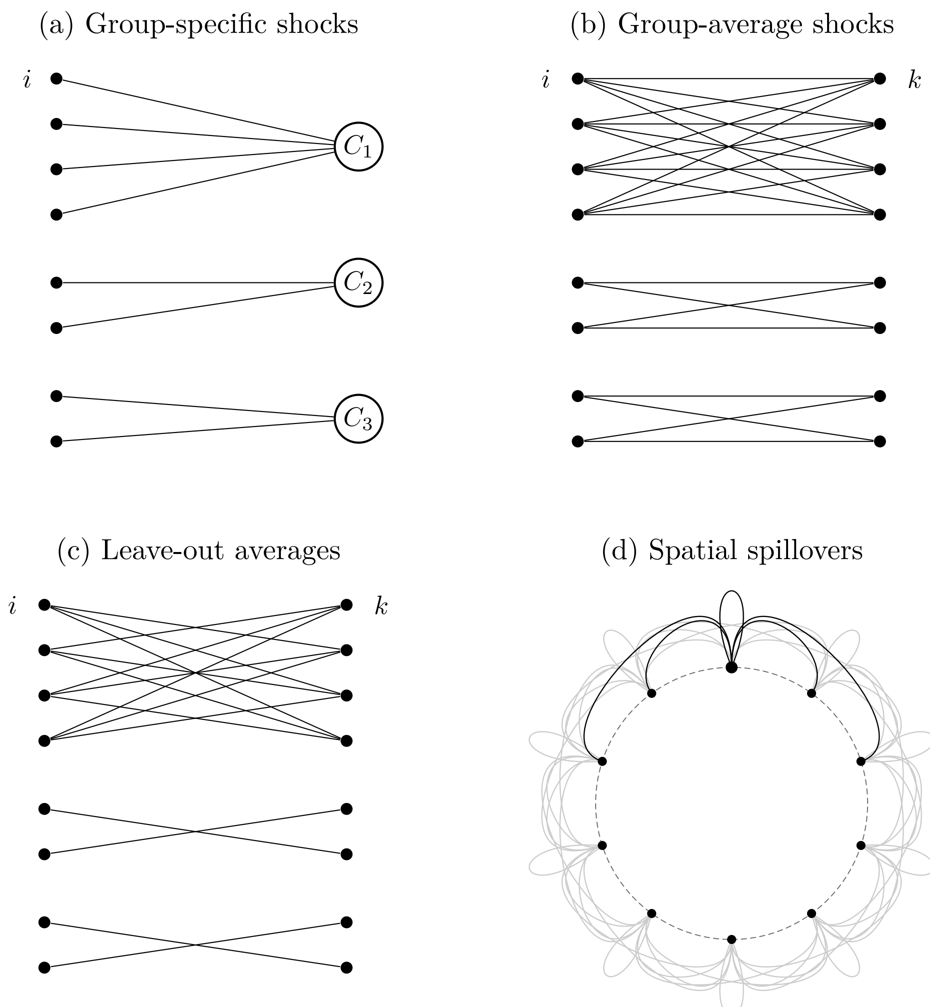
Group-Specific Shocks. This and remaining cases are visualized in Figure 2. Consider a simple bipartite graph in which observations are partitioned into J groups, $C_1, \ldots, C_J$, with shocks assigned at the group level: $K = J$ and $x_i = g_{j(i)}$ corresponding to $w_i = e_{j(i)}$, where e_j is the jth standard basis in $\mathbb{R}^K$ and $j(i) \in \{1, \ldots, K\}$ gives the group of observation i. Again Proposition 2 implies independent $g_j \overset{iid}{\sim}$ *Bernoulli*(0.5) is optimal. The optimal instrument is now proportional to $N_{j(i)}^{-1/(1+p)} \cdot (g_{j(i)} - 0.5)$ where $N_j = |C_j|$. Thus when $p = 1$ the estimator down-weights large groups by a factor of $1/\sqrt{N_{j(i)}}$ to ensure the random variation is equally spread out; not doing so would make the estimator noisy if the errors are very correlated within clusters. With higher p less of this reweighting is needed, with no reweighting in the limit $p = \infty$.

[14] Another problem with degenerate designs is that the asymptotic approximation that makes the approximate variance a useful concept may fail. Moreover, the estimator itself may not be well-defined with a probability that is not asymptotically small, as $\sum_i z_i(g) x_i = 0$ whenever $g = \mathbf{0}$; see footnote 10.

[15] While $\mathcal{F}_p(\sigma)$ allows for error dependence even across blocks, Proposition 2 implies that the same design would be optimal even if one were to *a priori* rule out cross-block dependence.

[16] Note that complete randomization, with exactly half of the units treated, is not optimal for $p < \infty$ as it induces a slight negative correlation across assignments which can be exploited by the adversary choosing the errors.

Figure 2: Graphs in Special Cases



Notes: This figure shows network graphs in the special cases considered in Section 2.3.

Group-Average Shocks. Now consider a case in which observations are partitioned into J groups $C_1, \dots, C_J$, but the shocks g_i are assigned to individuals ($K = N$) and fully propagate through the cluster-average: $x_i = \bar{g}_{j(i)}$ for $\bar{g}_j = \frac{1}{N_j}\sum_{i \in C_j} g_i$, which corresponds to $W = \operatorname{diag}(P_{N_1}, \dots, P_{N_J})$ for $P_n = \mathbf{1}_n \mathbf{1}_n'/n$. This is isomorphic to the previous case, with shocks that are perfectly correlated within groups and independent across groups being optimal: $g_i = \check{g}_{j(i)}$ with $\check{g}_j \overset{iid}{\sim} Bernoulli(0.5)$. The optimal instrument is again proportional to $N_{j(i)}^{-1/(1+p)}(\check{g}_{j(i)} - 0.5)$. Intuitively, there is no need to vary the shocks within groups since treatment is at the group level; perfect within-group shock correlation ensures the maximal variability of the spillover treatment while we still reweight by $N_{j(i)}^{-1/(1+p)}$ to more evenly spread out this variation.

Leave-Out Averages. Suppose in the previous case we instead have $N_j > 1$ for all j and $x_i = \bar{g}_{-i} = \frac{1}{N_{j(i)}-1}\sum_{k \in C_{j(i)}, k \neq i} g_k$. Appendix B.6 shows that an optimal design mixes between group-specific iid $Bernoulli(0.5)$ shocks with probability $\psi_j = (1 - r_j)/(1 + (N_j - 1)r_j)$ for $r_j = (N_j - 1)^{-2p}$, and independent individual-level $Bernoulli(0.5)$ shocks with probability $1 - \psi_j$, independently across groups.[17] This design is close to a more conventional saturation design (studied, e.g., in Baird et al. (2018)) with saturation rates of 0, 0.5, and 1, except with independent (rather than complete) randomization within groups. The optimal instrument is now proportional to $\left(1 + (N_{j(i)} - 1)\psi_{j(i)}\right)^{-1/(1+p)} \cdot \left((\bar{g}_{j(i)} - \frac{1}{2}) + (N_{j(i)} - 1)(\bar{g}_{j(i)} - g_i)\right)$. It is easy to verify that it is optimal to correlate the shocks more in larger groups: ψ_j increases in N_j, from $\psi_j = 0$ when $N_j = 2$ to $\psi_j \to 1$ when $N_j \to \infty$. Intuitively, in the smallest group of size $N_j = 2$, we are effectively back in the no-spillover case, while with $N_j \to \infty$ the leave-out average is equivalent to the group average from the previous case. For intermediate sizes, within-group correlation increases in p: as before, it is higher when less clustering is allowed in ε.

Spatial Spillovers. Finally, suppose the $N = K$ units are equal-spaced points on a circle, with treatment being an average of shocks to units within the distance of L: $x_i = \frac{1}{2L+1}\sum_{k=i-L}^{i+L} g_k$, where index summation is understood modulo N. Appendix B.7 shows that, for $p = 1$, the optimal design correlates shock assignments at the same spatial scale as the exposure mapping. Specifically, the correlation between assignments at circular distance d is $\max\{1 - d/(2L+1), 0\}$, linearly decaying in d. The design can be implemented by splitting the circle into blocks of $2L + 1$ consecutive units, randomly shifting this partition, and assigning shocks at the block level. The optimal instrument is $z^* = W^\dagger \tilde{g}$. Under the shifted-block implementation of the design, it takes values $(2L+1)\tilde{g}_i - 2L\bar{\tilde{g}}$ for units at the center of a block and $\bar{\tilde{g}}$ for all other units, where $\bar{\tilde{g}}$ is the average of all recentered shocks. Intuitively, the IV undoes the smoothing done by spatial averaging to mitigate the correlations that could otherwise be exploited by the adversary. These results can be extended to a two-dimensional grid, with x_i averaging the shocks within a certain distance in each coordinate.

[17] Note that this is just one way to induce the optimal $\operatorname{Var}_{\delta^*}[x]$. Other designs may deliver the same variance matrix and are therefore also optimal.

2.4 Multiple Exposures

We now consider a model in which the shocks g affect the outcomes through multiple linear exposure measures. For brevity, we consider two exposures but all results generalize immediately to three or more. For a known fixed $N \times K$ matrix $U = (u_i')_{i=1}^N$, we now have:

$$\begin{aligned} y_i &= \beta x_i + \tau h_i + \varepsilon_i, \\ x_i &= w_i' g, \qquad h_i = u_i' g. \end{aligned} \tag{4}$$

Our goal is still to minimize the worst-case approximate estimation variance of the effect of x_i.

While we consider a general formulation, this setting is especially relevant when the intervention and outcome units are the same and τ corresponds to the direct effect of the intervention:

$$y_i = \beta x_i + \tau g_i + \varepsilon_i \tag{5}$$

which corresponds to $U = I_N$.[18] In this context, there are three reasons why one would focus on the estimation power of β. First, the researcher may be most interested in the spillover effect, but still wants to acknowledge the presence of the direct effect. Second, even if direct and spillover effects are of equal interest, the spillover effect is usually more challenging to estimate. A design that yields sufficient power for it would usually deliver precise estimates of the direct effect, too. Finally, if the researcher is interested in the Global Average Treatment Effect (GATE), defined as the effect of switching from $g = \mathbf{0}_K$ to $g = \mathbf{1}_K$ via both direct and indirect effects, that problem can be reformulated as the one we solve here.[19]

To see how Theorem 1 and Propositions 1–2 generalize, collect the parameters and explanatory variables in $\theta = (\beta, \tau)'$ and $\boldsymbol{x}_i = (x_i, h_i)'$. Absent restrictions on how the error term can correlate with W and U, we again use recentered IV. For some recentered instrument vector $\boldsymbol{z}(\cdot) = (\boldsymbol{z}_i(\cdot))_{i=1}^N$, where $\boldsymbol{z}_i(\cdot) : \{0,1\}^K \to \mathbb{R}^2$ and $\mathbb{E}_\delta [\boldsymbol{z}_i(g)] = 0$, consider the estimator:

$$\hat{\theta}[\boldsymbol{z}] = \left(\sum_{i=1}^N \boldsymbol{z}_i(g) \boldsymbol{x}_i' \right)^{-1} \sum_{i=1}^N \boldsymbol{z}_i(g) y_i.$$

Define the approximate variance of $\hat{\beta}$ as:

$$\boldsymbol{\mathcal{V}}_{\delta,\mathcal{E}}[\boldsymbol{z}] = e_1' \mathbb{E}_\delta \left[\sum_{i=1}^N \boldsymbol{z}_i(g) \boldsymbol{x}_i' \right]^{-1} \mathrm{Var}_{\delta,\mathcal{E}} \left[\sum_{i=1}^N \boldsymbol{z}_i(g) \varepsilon_i \right] \mathbb{E}_\delta \left[\sum_{i=1}^N \boldsymbol{z}_i(g) \boldsymbol{x}_i' \right]^{-1\prime} e_1,$$

[18] Another class of applications is to multiplex networks (cf. Zenou (2026)).

[19] Specifically, the GATE equals $\gamma = \bar{w}\beta + \bar{u}\tau$ for known constants $\bar{w} = \frac{1}{N}\sum_{i,k} w_{ik}$ and $\bar{u} = \frac{1}{N}\sum_{i,k} u_{ik}$. Provided $\bar{w} \neq 0$, such that the indirect effect matters, one can then rewrite (4) as $y_i = \gamma \frac{x_i}{\bar{w}} + \tau \left(h_i - \frac{\bar{u}}{\bar{w}} x_i \right) + \varepsilon_i$ and apply our results to this reformulation, optimizing estimation power for γ while also including $h_i - \frac{\bar{u}}{\bar{w}} x_i$ in the model.

where $e_1 = (1,0)'$, with $\boldsymbol{\mathcal{V}}_{\delta,\mathcal{E}}[\boldsymbol{z}] = \infty$ whenever $\mathbb{E}_\delta\left[\sum_{i=1}^N \boldsymbol{z}_i(g)\boldsymbol{x}_i'\right]$ is singular. Also, as before, define

$$\mathcal{V}_{\delta,\mathcal{E}}[z] = \frac{\mathrm{Var}_{\delta,\mathcal{E}}\left[\sum_{i=1}^N z_i(g)\varepsilon_i\right]}{\mathbb{E}_\delta\left[\sum_{i=1}^N z_i(g)x_i\right]^2}.$$

An initial result shows that the problem of searching over recentered IV vectors $\boldsymbol{z}$ is equivalent to a constrained version of the original problem of searching over scalar z that requires z to be orthogonal to the "nuisance" exposure h:

Lemma 1. *Fix δ and $\mathcal{E}$. Let $\mathcal{Z}_\delta^{(2)}$ be the set of recentered $\boldsymbol{z}$ and suppose $\mathrm{tr}(\mathrm{Var}_\delta[h]) > 0$ for $h = (h_i)_{i=1}^N$. Then:*

$$\left\{\boldsymbol{\mathcal{V}}_{\delta,\mathcal{E}}[\boldsymbol{z}] : z \in \mathcal{Z}_\delta^{(2)}\right\} = \left\{\mathcal{V}_{\delta,\mathcal{E}}[z] : z \in \mathcal{Z}_\delta, \mathbb{E}_\delta\left[z'h\right] = 0\right\}.$$

That is, the set of attainable $\boldsymbol{\mathcal{V}}_{\delta,\mathcal{E}}[\boldsymbol{z}]$ equals the set of attainable $\mathcal{V}_{\delta,\mathcal{E}}[z]$ under the constraint $\mathbb{E}_\delta[z'h] = 0$. Moreover, this equivalence is constructive in both directions: any $\boldsymbol{z} \in \mathcal{Z}_\delta^{(2)}$ can be transformed into an instrument vector whose first component is some scalar $z \in \mathcal{Z}_\delta$ with $\mathbb{E}_\delta[z'h] = 0$, without changing the value of the objective, while any such scalar z can be implemented by the vector instrument $\boldsymbol{z}_i(g) = \big(z_i(g),\ h_i - \mathbb{E}_\delta[h_i]\big)$.

Imposing the orthogonality constraint yields the analog of Theorem 1:

Theorem 1′. *Fix δ and let*

$$\begin{aligned} &c_\delta \in \arg\min_{c\in\mathbb{R}} \mathrm{tr}\left(\left((W-cU)\Sigma_\delta(W-cU)'\right)^{p/(p+1)}\right), \\ &\tilde{x}_\delta^\perp = (W-c_\delta U)(g-\mathbb{E}_\delta[g]), \qquad S_\delta^\perp = (W-c_\delta U)\Sigma_\delta(W-c_\delta U)'. \end{aligned} \tag{6}$$

Suppose $S_\delta^\perp \neq 0$ and $\mathrm{tr}(U\Sigma_\delta U') > 0$. Then, for any $p \in [1,\infty]$ and $\sigma > 0$:

(a) The worst-case approximate variance with the optimal instrument is:

$$\min_{\boldsymbol{z}\in\mathcal{Z}_\delta^{(2)}} \max_{\mathcal{E}\in\mathcal{F}_p(\sigma)} \boldsymbol{\mathcal{V}}_{\delta,\mathcal{E}}[\boldsymbol{z}] = N^{1/p}\sigma^2\,\mathrm{tr}\left(\left(S_\delta^\perp\right)^{p/(p+1)}\right)^{-(p+1)/p}.$$

(b) Assume that either $p > 1$ or $p = 1$ and $\mathrm{rank}\Big((W-c_\delta U)\Sigma_\delta^{1/2}\Big) = \min\{N, \mathrm{rank}(\Sigma_\delta)\}$. Then the instrument $\boldsymbol{z}_i^*(g) = (z_\delta^*(g)_i, h_i - \mathbb{E}_\delta[h_i])$ minimizes $\max_{\mathcal{E}\in\mathcal{F}_p(\sigma)} \boldsymbol{\mathcal{V}}_{\delta,\mathcal{E}}[\boldsymbol{z}]$ over $\mathcal{Z}_\delta^{(2)}$, where:

$$z_\delta^*(g) = \left((S_\delta^\perp)^\dagger\right)^{1/(p+1)} \tilde{x}_\delta^\perp.$$

The key difference from the optimal instrument in the single-exposure case is that the matrix W is replaced with an appropriate residualization $W - c_\delta U$ where c_δ (which plays the role of a Lagrange multiplier on the $\mathbb{E}_\delta[z'h] = 0$ constraint) can be solved for by the scalar optimization

(6). The additional technical assumption for $p = 1$ is imposed because the objective (6) can be non-differentiable, which may lead to a different analytical form of $z^*_\delta(g)$. To avoid this issue, we require that the residualized exposure matrix retains as much assignment-induced variation as its dimensions allow. Given orthogonality between the optimal IV for x_i and the exposure h_i, the IV for h_i is simply $h_i - \mathbb{E}_\delta[h_i]$.

Analogs of Propositions 1–2 also follow:

Proposition 1′. *In the Theorem 1′ setting, suppose there is* $\delta \in \mathcal{D}$ *such that* $\operatorname{tr}(U\Sigma_\delta U') > 0$ *and* $\min_{c\in\mathbb{R}} \operatorname{tr}\left(((W - cU)\Sigma_\delta(W - cU)')^{p/(p+1)}\right) > 0$*. Then*

$$\delta^*, \boldsymbol{z}^* \in \arg \min_{\delta\in\mathcal{D}, \boldsymbol{z}\in\mathcal{Z}^{(2)}_\delta} \max_{\mathcal{E}\in\mathcal{F}_p(\sigma)} \boldsymbol{\mathcal{V}}_{\delta,\mathcal{E}}[\boldsymbol{z}]$$

is solved by

$$\delta^* \in \arg\max_{\delta\in\mathcal{D}} \operatorname{tr}\left((S^\perp_\delta)^{p/(p+1)}\right).$$

The optimal instrument is $\boldsymbol{z}^* = \boldsymbol{z}^*_{\delta^*}$ *from Theorem 1′, as long as, for* $p = 1$, $\operatorname{rank}\left((W - c_\delta U)\Sigma^{1/2}_\delta\right) = \min\{N, \operatorname{rank}(\Sigma_\delta)\}$ *.*

Proposition 2′. *There is a* δ^* *from Proposition 1′ with* $\mathbb{E}_{\delta^*}[g_k] = 0.5$ *for all* k*. Further, if* W *and* U *are jointly block diagonal under the same row-column partition, there is such a* δ^* *with independent subvectors of* g *across blocks.*

We note that, unlike with Proposition 2, the optimal design cannot generally be solved block-by-block in the block-diagonal case because c_δ is common across all blocks.

We illustrate these results in a simple but nontrivial special case: when x_i is the leave-out average of shocks in the cluster to which i belongs and $h_i = g_i$ captures the direct effect. Appendix B.6 shows that, at least when N_j is even, it is optimal to mix between cluster-specific *Bernoulli*(0.5) shocks and complete randomization of the shocks within each cluster, independently across clusters. In the special case where all clusters are of the same size, $N_j = n$, one can obtain a solution more similar to the one from Section 2.3: mix between cluster-specific shocks with probability $\psi^\perp = (1 - r^\perp)/(1 + (n-1)r^\perp)$ for $r^\perp = (n-1)^{-2p/(2p-1)}$ and independent *Bernoulli*(0.5) randomization with probability $1 - \psi^\perp$. The optimal spillover instrument is proportional to $(\bar{g}_{j(i)} - 0.5) + (n - 1)^{1/(2p-1)}(\bar{g}_{j(i)} - g_i)$. Relative to the case without direct effects, there is generally more randomization within clusters ($\psi^\perp \le \psi$, with strict inequality for $n > 2$ and $p > 1$), which helps isolate spillover effects from the direct effects. The exception is when $p = 1$: with both homogeneous and heterogeneous N_j, the design and instrument are then the same as in the original problem, because the objective already creates enough variation to identify direct effects.

2.5 Other Extensions

We derive three further extensions to the baseline results. First, Appendix B.8 shows how the results extend when the researcher includes an intercept and possibly other predetermined covariates r_i in estimation. We formalize this idea by enriching the class of error distributions to $\varepsilon_i = r_i'\gamma + \tilde{\varepsilon}_i$ where $\tilde{\varepsilon} = (\tilde{\varepsilon}_i)$ satisfies $\mathbb{E}_{\mathcal{E}}[\tilde{\varepsilon}\tilde{\varepsilon}'] \in \mathcal{F}_p(\sigma)$ while γ is unrestricted. The results in Theorem 1 and Propositions 1–2 continue to hold, with the exposure matrix W residualized columnwise on the covariates.

Second, Appendix B.9 discusses how Theorem 1 and Proposition 1 extend to the general formula treatment setting where the x_i are nonlinear (but still known) i-specific formulas combining W and g. This allows, for example, x_i to be an indicator for i having at least one treated friend in their social network or the more elaborate formulas considered in Borusyak and Hull (2023). The proofs to Theorem 1 and Proposition 1 turn out to extend verbatim after defining the general $\tilde{x}_\delta = x - \mathbb{E}_\delta[x]$ and $S_\delta = \mathrm{Var}_\delta[x]$, making them a robust characterization of optimal design with formula treatments. Proposition 2, however, does not extend cleanly: it is no longer without loss of generality to consider designs with 0.5 marginals. The following discussion of feasible design computation and inference is also reliant on linearity of the spillover treatment formula.

Finally, Appendix B.10 considers settings where the researcher faces a constraint of a marginal treatment probability $q \neq 0.5$ (common across units), e.g. due to a fixed budget or government target. Theorem 1 applies to any design and is unaffected by the constraints on feasible designs. We therefore show that Proposition 1 extends naturally with such constraints, as well as the second claim of Proposition 2 and the results on computation, below.

3 Feasible Designs and Inference

We now introduce a relaxed version of the optimal design problem. This relaxation admits a closed-form solution in special cases and a computationally efficient approximation in general. Moreover, this approximation can yield $\sqrt{K}$-consistent recentered IV estimators with $p < \infty$, as well as a simple asymptotic inference procedure.

3.1 The Relaxed Problem

Solving the Proposition 1 problem is computationally intractable for even moderate K. While its objective depends on the experimental design δ only through the $K \times K$ covariance matrix $\mathrm{Var}_\delta[g]$, the class of all $K \times K$ covariance matrices achievable with binary treatments, $\mathcal{C}_K = \{\mathrm{Var}_\delta[g] : \delta \in \mathcal{D}\}$, is high-dimensional. The computation time required to find an optimal $\mathrm{Var}_\delta[g]$, by searching over convex combinations of the 2^K possible treatment vectors in $\{0,1\}^K$, is generally exponential in K.

To make progress, we follow Goemans and Williamson (1995) and Thiyageswaran et al. (2026) in considering a relaxed problem with a polynomial-time solution in K. Specifically, we replace the optimization over $\mathcal{C}_K$ with optimization over $\mathcal{Q}_K = \left\{\Sigma \in \mathbb{S}^K : \Sigma \succeq 0, \Sigma_{kk} = 1/4 \text{ for all } k\right\}$, where $\mathbb{S}^K$ is the set of $K \times K$ real and symmetric matrices. Thus $\mathcal{Q}_K$ is the set of $K \times K$ covariance matrices with 1/4 on the diagonal, containing $\mathcal{C}_K$ when we restrict attention to designs with $\mathbb{E}_\delta[g_k] = 0.5$ (without loss of generality, by Proposition 2). The relaxed Proposition 1 problem is then:

$$\max_{\Sigma \in \mathcal{Q}_K} \operatorname{tr}\left((W \Sigma W')^{\frac{p}{p+1}} \right). \tag{7}$$

Searching over $\mathcal{Q}_K$ is feasible for moderate K using an interior point solver for conic optimization problems. Two additional results further improve computational efficiency without any additional approximation cost. First, if W can be split into independent blocks $W_{(1)}, \ldots, W_{(B)}$ with $K(b)$ shocks in each, we can use Proposition 2 to solve the optimal design problem separately by block. Second, as the following result shows, we can characterize the solution explicitly in terms of a lower-dimensional convex optimization problem over the $K-1$-dimensional simplex $\Delta_K = \left\{ \ell \in \mathbb{R}^K : \ell_k \geq 0, \sum_{k=1}^K \ell_k = 1 \right\}$ rather than $\mathcal{Q}_K$; this problem is amenable to gradient-based optimization approaches that scale more efficiently with K.

Proposition 3. *For $p < \infty$, let*

$$\Sigma^* = \frac{1}{4} D_{\ell^*}^{-1/2} \frac{M_{\ell^*}^p}{\operatorname{tr}\left(M_{\ell^*}^p \right)} D_{\ell^*}^{-1/2},$$

where, for $\ell > 0$,

$$D_\ell = \operatorname{diag}(\ell_1, \ldots, \ell_K), \qquad M_\ell = D_\ell^{-1/2} W' W D_\ell^{-1/2},$$

and

$$\ell^* \in \arg\min_{\ell \in \Delta_K} \operatorname{tr}\left(M_\ell^p \right), \tag{8}$$

with $\ell^* > 0$. Then Σ^* solves the relaxed problem (7).

This characterization is especially helpful in two special cases where it yields closed-form solutions of the relaxed problem:

Corollary 1. *For $p = 1$, the relaxed problem (7) is solved by:*

$$\Sigma^*_{kl} = \frac{1}{4} \frac{w_{\cdot k}' w_{\cdot l}}{\|w_{\cdot k}\|_2 \|w_{\cdot l}\|_2}, \qquad k, l \in \{1, \ldots, K\},$$

where $w_{\cdot k}$ is the kth column of W. That is, in the relaxed problem's solution, the correlation of assignments of any two shocks equals the cosine similarity of the vectors of exposures to those shocks.

Corollary 2. *For $p < \infty$, if all diagonal elements of $(W'W)^p$ are equal, the relaxed problem (7) is solved by:*

$$\Sigma^* = \frac{1}{4} \frac{(W'W)^p}{\operatorname{tr}\left((W'W)^p \right) / K}.$$

The Theorem 1 optimal IV corresponding to Σ^* is $z^* = a \cdot (W')^\dagger \tilde{g}$ which depends on p only through $a = \left(4 \operatorname{tr}\left((W'W)^p \right) / K \right)^{1/(p+1)}$. Moreover, when W consists of multiple independent blocks with the diagonal elements of $(W'W)^p$ equal within blocks, these expressions apply block by block, with a possibly varying across blocks.

Corollary 1 captures a simple intuition: shocks to two intervention units should be correlated to the extent that they are connected, in the sense that the same outcome units are exposed to them. This turns out to be the optimal solution to the relaxed problem when $p = 1$. Corollary 2 extends this result to arbitrary p, as long as each block of W is sufficiently symmetric—a restrictive condition that nevertheless includes all of the special cases in Section 2.3. The corresponding solution is particularly intuitive when W is nonnegative and p is an integer. In that case, the matrix $(W'W)^p$ captures pth-degree connections between intervention units k. When $p = 2$, for example, the solution suggests correlating two shocks when they are either connected directly or connected to the same third shock. The worst-case loss from such longer-range shock correlations is lower with higher p, as error correlations are more restricted. As $p \to \infty$, this solution correlates all shocks in the same connected component of the bipartite network defined by W.

Corollary 2 further characterizes the IV that is optimal if the relaxed problem's solution can be implemented (an issue we discuss below). This optimal IV satisfies $W'z^* \propto \tilde{g}$, meaning that partial whitening entails the inverse operation to the exposure mapping, $\tilde{x} = W\tilde{g}$. For instance, when x involves some averaging of the (correlated) shocks, z^* performs their deconvolution.[20]

All of these results generalize immediately to settings with an intercept and possibly other predetermined covariates in estimation, by first residualizing W on those covariates column by column. For instance, for $p = 1$ and when estimation includes an intercept as the only covariate, Corollary 1 applied to the demeaned $w_{\cdot k}$ implies that shock correlations under the relaxed problem's solution are equal to the correlations of the $N \times 1$ exposures to those shocks (rather than cosine similarities without the intercept). This solution positively correlates shocks with similar exposure while introducing slight negative correlations between shocks with non-overlapping exposure.[21]

We can also generalize the relaxed problem to multiple exposures, where it remains computationally efficient. By the same arguments as in the Proposition 1′ proof, the relaxed problem is:

$$\min_{c \in \mathbb{R}} \max_{\Sigma \in \mathcal{Q}_K} \operatorname{tr}\left(\left((W - cU)\Sigma(W - cU)'\right)^{\frac{p}{p+1}} \right). \tag{9}$$

Here the inner problem has the same structure as before with a single exposure, with W replaced by $W - cU$; the computationally efficient form of Proposition 3 and its closed-form special cases thus apply.[22] Additionally, when W and U are jointly block-diagonal, the inner problem can be split block-by-block for any c. The outer problem is then a scalar optimization, amenable to a golden section search over a compact subset of $\mathbb{R}$.[23]

[20]Interestingly, the optimal IV is invariant to p, up to the proportionality constant a. With larger p, the shocks are more strongly correlated; thus, the same IV that fully whitens the shocks when $p = 1$ does only partial whitening for larger p. We also note that when W consists of independent blocks the constant a varies across blocks; thus, p affects the reweighting of blocks but not the optimal IV within blocks.

[21]That is, if $w_{\cdot k}, w_{\cdot l} \geq 0$ and $w_{\cdot k}' w_{\cdot l} = 0$ (no exposure overlap), Σ^*_{kl} is proportional to the sample correlation of $w_{\cdot k}$ and $w_{\cdot l}$, which is negative.

[22]The Proposition 3 problem becomes $\min_{c \in \mathbb{R}} \min_{\ell \in \operatorname{int}(\Delta_K)} \operatorname{tr}\left\{ \left(D_\ell^{-1/2} (W - cU)' (W - cU) D_\ell^{-1/2} \right)^p \right\}$.

[23]The minimizer is guaranteed to exist in a compact subset of $\mathbb{R}$ as long as $\max_{\Sigma \in \mathcal{Q}_{\mathcal{K}}} \operatorname{tr}\left(\left((W - cU)\Sigma(W - cU)'\right)^{\frac{p}{p+1}} \right) \geq \operatorname{tr}\left(\left((W - cU)I/4(W - cU)'\right)^{\frac{p}{p+1}} \right) \to \infty$ when $|c| \to \infty$, which requires only that $U \neq 0$.

3.2 Feasible Implementation

Two challenges remain once a solution Σ^* to the relaxed problems in Equation (7) or (9) is found. First, this Σ^* may not correspond to an element of $\mathcal{C}_K$: i.e., it may not be implementable via binary shock vectors. Second, even if it is theoretically implementable, there is no known computationally feasible procedure for sampling binary vectors from a generic covariance matrix. These problems can be overcome in simpler special cases: the designs in Section 2.3 are indeed implementations of the relaxed problem solution.

Outside special cases, we address both challenges by using a simple Gaussian rounding procedure to sample binary assignments with the variance matrix approximating Σ^*, following Goemans and Williamson (1995). Specifically, we draw $\xi \sim \mathcal{N}(0, \Sigma^*)$ and set $g_k^{GR} = \mathbf{1}\left[\xi_k \geq 0\right]$.[24] This procedure distorts correlations in a known way: $\text{Corr}\left[g_k^{GR}, g_l^{GR}\right] = \arcsin\left[\text{Corr}\left[\xi_k, \xi_l\right]\right] / \frac{\pi}{2} \neq \text{Corr}\left[\xi_k, \xi_l\right]$; see Goemans and Williamson (1995) and the illustration in Appendix Figure A1. Given this Gaussian rounding design, we find the optimal IV based on $S_{GR} = W\text{Var}\left[g^{GR}\right] W'$ for

$$\Sigma^{GR} = \text{Var}\left[g^{GR}\right] = \frac{1}{2\pi} \arcsin\left[4\Sigma^*\right], \tag{10}$$

with $\arcsin\left[\cdot\right]$ applied entrywise. Algorithm 1 summarizes the entire procedure.

Beyond simplicity in implementation, this procedure offers two advantages. First, Appendix B.11 shows the cost of using this relaxed and rounded version of the problem is bounded: the worst-case approximate variance of the resulting estimator is at most $\pi/2$ times that of the original problem.[25] Second, as we next show, the fact that our shocks are a simple transformation of a correlated Gaussian vector is helpful for establishing a central limit theorem for the recentered IV estimator and asymptotically valid inference, even when all errors are mutually correlated.

3.3 Asymptotic Normality and Inference

We now consider the asymptotic behavior of the estimator $\hat{\beta} = \hat{\beta}\left[z^*\right]$ corresponding to the Gaussian rounding implementation of the Section 3.1 relaxed-optimal design, with the Theorem 1 optimal IV z^*, as in Algorithm 1. We show that choosing a small integer p can yield sparsity of the implied shock covariance matrix, which in turn can yield a central limit theorem for $\hat{\beta}$ and a simple inference procedure.

Let $d(\Sigma) = \max_k \sum_{l=1}^K \mathbf{1}\left\{\Sigma_{kl} \neq 0\right\}$ measure the sparsity of a covariance matrix Σ by the maximum number of non-zero covariances across all rows. We consider a sequence of data-generating processes indexed by K. For $K \to \infty$ (which under Assumption 1 will require $N \to \infty$ as well) we assume:

[24] This procedure simplifies further in the special case of Corollary 1: draw $\eta \sim \mathcal{N}(0, I_N)$ and set $g_k^{GR} = \mathbf{1}\left[(W'\eta)_k \geq 0\right]$. Indeed, we can write $g_k^{GR} = \mathbf{1}\left[\xi_k \geq 0\right]$ for $\xi_k = \sum_i w_{ik}\eta_i / 2\left\|w_{\cdot k}\right\|_2$, where $(\xi_1, \dots, \xi_K) \sim \mathcal{N}(0, \Sigma^*)$ for Σ^* from Corollary 1.

[25] We find much smaller approximation errors in applications: for all designs in Section 4, the worst-case approximate variance exceeds the variance of the original problem by no more than 6%. This is computed by comparing the worst-case variance arising from the Gaussian rounded design to that of the unrounded solution to the relaxed problem (7), which is a lower bound for the variance of the optimal solution to the original problem.

Algorithm 1 Feasible Optimal Experimental Design and IV

0. **Input:**
   - $N \times K$ matrix of the indirect exposure of interest, W
   - *Optional:* $N \times K$ matrix of the other included exposure, U; e.g., $U = I$ for direct exposure
   - *Optional:* $N \times L$ matrix of predetermined covariates, R
   - Tuning parameter $p \in [1, \infty)$. Set lower p for higher robustness to non-spherical errors
1. If $R \neq \emptyset$, replace W and U with their columnwise residuals after projecting on R
2. Split (W, U) jointly into independent blocks of rows and columns, $(W_{(b)}, U_{(b)})$, if any
   - *Optional:* to ease computation, at the cost of some approximation error, also split approximately independent blocks (e.g., if $R = \mathbf{1}$, the zero elements in the original (W, U) for rows and columns from different blocks will be replaced in step 1 with $-1/N \approx 0$)
3. Fix $c = 0$. Solve for the relaxed-optimal shock covariance matrix Σ^* block-by-block:
   - If $p = 1$, use the Corollary 1 closed-form solution
   - If all diagonal elements of $(W'_{(b)} W_{(b)})^p$ are equal, use the Corollary 2 closed-form solution
   - Else, use the Proposition 3 numerical optimization
4. If $U \neq \emptyset$, choose c_δ that solves the optimization problem in (9) by repeating Step 3 for $c \neq 0$ and $W - cU$ replacing W (on a grid or using gradient-based methods)
5. Randomize shocks using the Gaussian rounding design δ from Section 3.2. **Output** g
6. Using $\mathrm{Var}_\delta [g] = \Sigma^{GR}$ from (10), choose the recentered instrument:
   - If $U = \emptyset$, follow Theorem 1. **Output** $z^*_\delta(g)$
   - If $U \neq \emptyset$, follow Theorem 1′. **Output** $\boldsymbol{z}^*_\delta(g)$

Assumption 1. *There are constants* $\underline{\lambda} > 0$, $\bar{\lambda} \in (0,\infty)$, $\bar{d} < \infty$, $h^* \in (0,\infty)$, *and* $v^* \in (0,\infty)$ *such that:*

(a) $d(\Sigma^*) \leq \bar{d}$ and $\lambda_{\max}(W'W) \leq \bar{\lambda}$;

(b) $\lambda^+_{\min}(S_{GR}) > \underline{\lambda}$ *with* $\lambda^+_{\min}(\cdot)$ *denoting the smallest positive eigenvalue;*

(c) $h_K = K^{-1}\operatorname{tr}(S_{GR}^{p/(p+1)}) \xrightarrow{p} h^*$*;*

(d) $v_K = K^{-1}\varepsilon' S_{GR}^{(p-1)/(p+1)}\varepsilon \xrightarrow{p} v^*$*;*

(e) $\frac{1}{K^{3/2}}\sum_{k=1}^K |b_k|^3 = o_p(1)$ *for* $b = W'(S^\dagger_{GR})^{1/(p+1)}\varepsilon$,

where statements (a) and (b) hold with probability $1 - o(1)$ *and all probability statements are with respect to the joint distribution of* (W, ε, ξ).

The first part of Assumption 1(a) is the key sparsity condition imposed on the solution to the relaxed problem in Equation (7). While the sparsity of Σ^* is formally an asymptotic condition, the researcher can heuristically check whether $d(\Sigma^*)$ is small compared to K; at the cost of more complex primitive conditions, we can accommodate $d(\Sigma^*)$ growing slowly in K.[26] Since $\frac{1}{N}\|x\|_2^2 = \frac{1}{N}\|Wg\|_2^2 \leq \lambda_{\max}(W'W)\cdot\frac{1}{K}\|g\|^2\cdot\frac{K}{N}$, the second part precludes x from diverging provided $K \asymp N$. Appendix A.6 provides sufficient conditions for both parts of this assumption. It shows that $d(\Sigma^*)$ is upper-bounded by $(d(W'W))^p$ under the assumptions of Proposition 3. With $p = \infty$, no sparsity of Σ^* is guaranteed within any connected component of the graph, and so designs that are optimal for $p = \infty$ may not yield consistent $\hat{\beta}$ when W does not consist of many independent blocks and when errors are also strongly dependent. But for smaller integer p, the optimal design only correlates units that are connected by paths in $W'W$ of lengths p or less. Thus, for finite integer p, both parts of Assumption 1(a) are guaranteed when maximum row and column degree of W as well as the maximum absolute value of its elements are all bounded. This set of conditions in turn requires $K \asymp N$ in non-degenerate cases, as also shown in Appendix A.6.

In addition to sparsity, Assumption 1 imposes regularity conditions to ensure the joint distribution of spillover treatments x_i and errors ε_i is well-behaved. Assumption 1(b) precludes near-collinearity of exposures to ensure that the whitening matrix remains well-conditioned. Yet, this condition allows S_{GR} to be degenerate, which happens, e.g., when some observations have the same exposure to all shocks. Noting that $h_K = \mathbb{E}_\delta\left[\frac{1}{K}x'z^* \mid W\right]$, Assumption 1(c) implies that W and the experimental design are such that there is sufficient variation in the spillover treatments x_i so that the first stage is not degenerate asymptotically. Noting further that $v_K = \frac{1}{K}\mathrm{Var}_\delta\left[\varepsilon' z^* \mid W, \varepsilon\right]$, Assumption 1(d) ensures that the finite-sample variance of $\varepsilon' z^*$ conditional on W and ε (scaled appropriately) converges to a deterministic constant. This is a law of large numbers for the quadratic

[26]Theorem 2 can be extended to cases where Σ^* is approximately sparse, in that it contains many small entries but has row and column sums that are bounded or grow very slowly. This extension is particularly useful when predetermined covariates are included, since then the effective W and corresponding Σ^* will not generally be sparse. In such cases we could allow the operator norm of Σ^* to grow slowly by a smoothing argument and the second order Poincaré inequality of Chatterjee (2009), as applied in Lei et al. (2018). We thank Lihua Lei for pointing this out.

form v_K and it imposes regularity on the joint distribution of W and ε. For example, conditional on a sequence of W, if ε is drawn from a mixture of two deterministic sequences that lead to different limits for v_K, then this assumption is violated; when dependence in ε is sufficiently weak or local, then this assumption is satisfied. Assumption 1(e) is an anti-concentration condition for the implicit weights of the numerator $\frac{1}{K}\varepsilon' z^*$; combined with the sparsity condition, it ensures that the numerator can be asymptotically approximated by sums of independent components.

Under these conditions, $\widehat{\beta}$ is $\sqrt{K}$-consistent:

Theorem 2. *Under Assumption 1 and the Gaussian rounding design based on the relaxed problem's solution from Proposition 3, the Theorem 1 optimal recentered IV estimator $\hat{\beta}$ satisfies*

$$\sqrt{K}(\widehat{\beta} - \beta) \Rightarrow \mathcal{N}\left(0, \frac{v^*}{h^{*2}}\right).$$

Asymptotically valid inference is straightforward from this result using the normal approximation and a simple plug-in estimator for the standard error, $\sqrt{\hat{v}/(Kh_K^2)}$. Here S_{GR} and therefore h_K are known. Moreover, the plug-in estimator for v^* is straightforward to compute: $\hat{v} = \frac{1}{K}\hat{\varepsilon}' S_{GR}^{(p-1)/(p+1)}\hat{\varepsilon}$ for $\hat{\varepsilon} = y - \hat{\beta}x$. The convergence in probability of $\hat{v}$ to v^* follows from consistency of $\widehat{\beta}$ and Assumption 1 (see Appendix A.6).

4 Applications

4.1 Bipartite Experiment: Cai et al. (2015)

Setup. Our first application is the bipartite experiment of Cai et al. (2015) who study how social networks affect the adoption of weather insurance by randomly assigning rice farmers in villages in China to simple vs. intensive information sessions over two rounds. They collect information on demographics, friendship, and insurance take-up for each farmer. We calibrate a simulation to a simplified version of the social network effect specification in Table 2, Column 2 of their paper. For the sample of farmers assigned to either the simple or intensive information session in the second round, it examines how insurance take-up varies with the fraction of friends that are treated in the first round. Specifically, our outcome units i are those randomized in the second round ($N = 1,274$),[27] and our intervention units k are those randomized in the first round with at least one friend in the second round ($K = 995$). As in Cai et al. (2015), we define $x_i = \sum_k w_{ik} g_k$ where w_{ik} equals an indicator that i and k are friends divided by the total number of friends i has (including friends outside the first-round sample). Thus, x_i measures the percentage of the second round unit's friends who were treated in the first round. Each i and k is assigned to one of 44 administrative villages, which are administrative units that each combine multiple local communities; throughout this analysis we drop 10 of 6,251 friendships that are across administrative villages so W has block-diagonal structure that eases computational burden. The top row of Figure 3(a) shows the exposure

[27] More precisely, outcome units are those used to estimate spillover effects on insurance take-up (from "Type 1" villages) rather than price effects.

matrix W for one example village, along with the $W'W$ matrix that captures similarity in exposure weights across intervention units.

We estimate the specification $y_i = \alpha + \beta x_i + \upsilon_i$ using recentered IV without whitening. Our simple specification has a smaller estimated coefficient of $\beta^* = 0.1408$ compared to the specification in Cai et al. (2015), which uses OLS as an estimator and includes an additional set of controls and village fixed effects, rather than recentering.

For $S = 1000$ simulations, we generate the outcome as $y_i = \beta^* x_i + \varepsilon_i$, where g_k varies by the experimental design, W is fixed, and we draw $\varepsilon = (\varepsilon_i)$ from six different data-generating processes (DGPs), i.e., joint distributions that vary the structure of the $N \times N$ matrix $\mathbb{E}_{\mathcal{E}}[\varepsilon\varepsilon']$.[28] For the first five DGPs, the distribution of the errors is calibrated so that the standard error of the unwhitened recentered IV estimator of β under a 50% Bernoulli RCT matches the standard error of the estimated coefficient in the data. The error distributions for the simulation are as follows:

- Homoskedastic: $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \nu^2)$, with $\nu = 0.56$;
- Heteroskedastic: $\varepsilon_i \sim \mathcal{N}(0, \nu^2(1 + \gamma d_i))$, independent across units i, where d_i is the number of i's first-round friends, $\nu = 0.34$ and $\gamma = 0.82$, which implies 50% of the variance is driven by the heteroskedastic component;
- Degree-in-Mean: $\varepsilon_i = \gamma d_i + \nu\zeta_i$, where $\zeta_i \overset{iid}{\sim} \mathcal{N}(0, 1)$, $\gamma = 0.29$, and $\nu = 0.31$, with γ calibrated such that 50% of the cross-sectional variation of ε_i arises from the mean;
- Village-Correlated: $\varepsilon_i = \nu \cdot (\zeta_i + u_{v(i)})$, where $\nu = 0.33$, $v(i)$ is the village of individual i, and $u_v \overset{iid}{\sim} \mathcal{N}(0, 1)$, $\zeta_i \overset{iid}{\sim} \mathcal{N}(0, 1)$;
- Network-Correlated: $\varepsilon = \nu\zeta + \gamma W_{\text{raw}}\eta$, where $\zeta_i \overset{iid}{\sim} \mathcal{N}(0, 1)$, $\eta_k \overset{iid}{\sim} \mathcal{N}(0, 1)$, and W_{raw} is the un-normalized friendship matrix, so $W_{\text{raw},i,k} = 1$ if i and k are friends and is otherwise 0. The parameters $\gamma = 0.29$ and $\nu = 0.18$ are calibrated so that 50% of the variance comes from the network component;
- Estimated Residuals: $\varepsilon_i = \hat{\upsilon}_i$, the residuals from the regression on the real data.

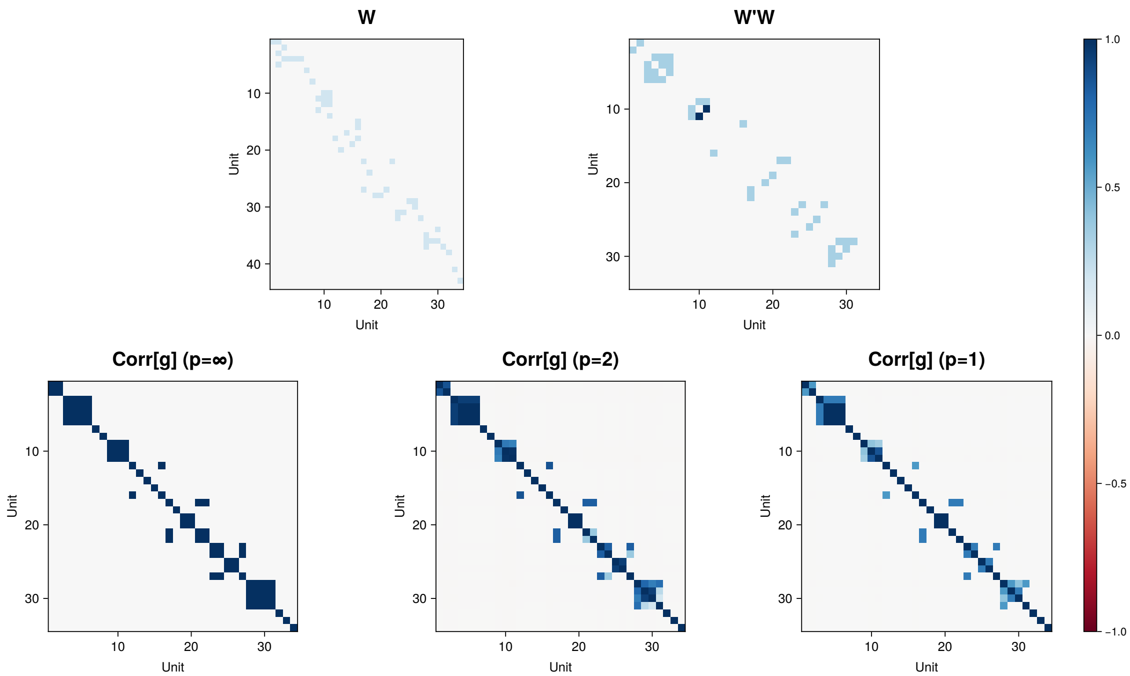
Results. Following Algorithm 1, we solve for the feasible optimal design and recentered IV estimator of the indirect effect for several choices of p. The bottom row of Figure 3(a) plots a correlation heatmap for these designs. As our theory suggests, $p = \infty$ involves perfectly correlating all treatments within clusters of connected farmers (with independent randomization across villages), while lower p implies more diffuse correlation structures.[29]

[28] While the original outcomes in Cai et al. (2015) are binary, our generated outcomes are continuous.

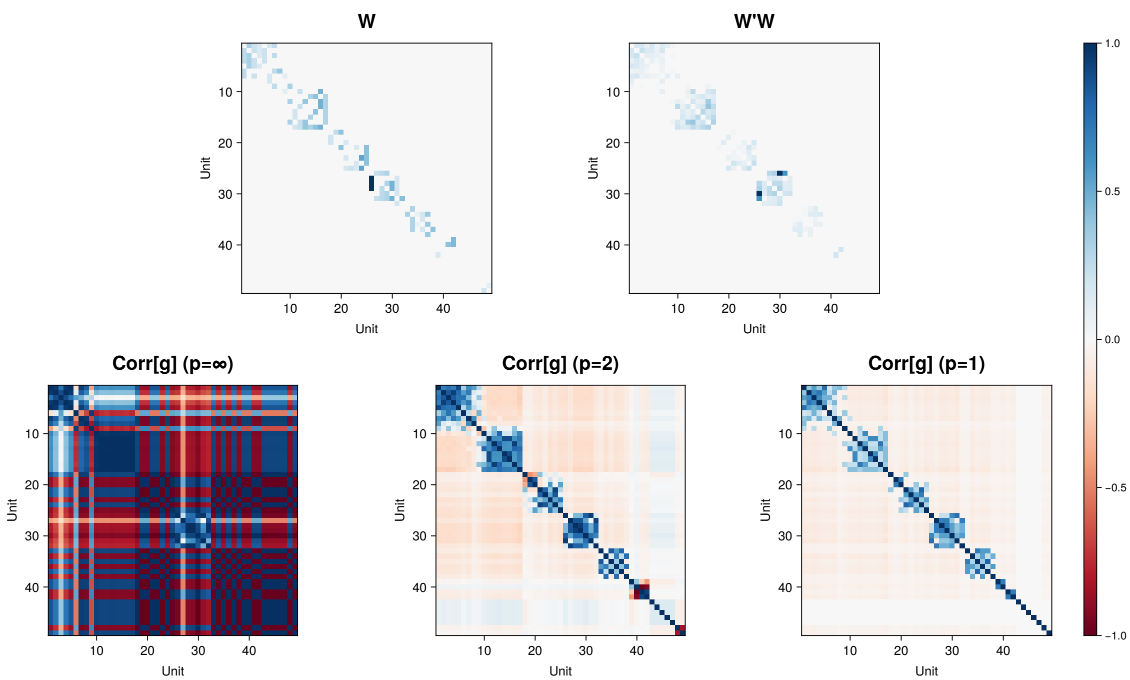
[29] Appendix Figure A2 shows how the optimal designs in both applications trade off isotropy, as measured by $\text{tr}(S_{\delta^*})/\lambda_{\max}(S_{\delta^*})$, and signal $\text{tr}(S_{\delta^*})$. As expected, the frontier moves down and to the right as $p \to \infty$. Notably, the baseline independent Bernoulli design (indicated by the red RCT dot) has higher isotropy than any optimal design but also much lower signal.

Figure 3: Spillover Structure and Optimal Designs in Applications

(a) An example village in Cai et al. (2015), $K = 34$, $N = 44$



(b) Miguel and Kremer (2004), $K = N = 49$



Notes: Panel (a) shows an example of one village (Xinlian) in the Cai et al. (2015) data. The first row shows the exposure matrix W and the matrix $W'W$ with the kl element measuring the extent to which k and l are friends of the same farmers i. The second row shows the correlation matrices of the Algorithm 1 feasible optimal designs corresponding to $p = 2$ and $p = 1$, along with the analogous solutions for $p = \infty$. Panel (b) reports the same objects for the full set of schools in Miguel and Kremer (2004), except W and $W'W$ are normalized to be between 0 and 1 by dividing each matrix element-wise by its maximum value.

Table 1: RMSE and Effective Sample Size Gains for the Cai et al. (2015) Application

Design and Estimator:	Homo-skedastic (1)	Hetero-skedastic (2)	Deg.-in-Mean (3)	Village-Corr. (4)	Network-Corr. (5)	Estimated Residuals (6)
RCT Benchmark	0.150 [+0%]	0.139 [+0%]	0.142 [+0%]	0.138 [+0%]	0.141 [+0%]	0.129 [+0%]
Optimal, $p = 1$	0.131 [+32%]	0.123 [+27%]	0.121 [+37%]	0.115 [+44%]	0.107 [+73%]	0.111 [+35%]
Optimal, $p = 2$	0.115 [+70%]	0.111 [+57%]	0.119 [+41%]	0.105 [+72%]	0.106 [+75%]	0.103 [+55%]
Optimal, $p = \infty$	0.095 [+151%]	0.099 [+96%]	0.133 [+13%]	0.117 [+39%]	0.107 [+73%]	0.099 [+71%]

Notes: For the Cai et al. (2015) application, this table reports the root mean-squared error of different design-estimator pairs (rows) under different data-generating processes for the errors (columns). The first row is the baseline of independent Bernoulli assignment and unwhitened recentered IV. The other rows are the Algorithm 1 feasible optimal design and optimal IV for different finite values of p (and the analogous solutions for $p = \infty$), with $R = \mathbf{1}$. The text describes the six error DGPs. Increases in effective sample sizes, given in brackets, are computed as the squared ratio of inverse RMSE under the focal scenario relative to the RCT baseline with the same error distributions, minus one.

Table 1 reports root mean-squared error (RMSE) for the optimal recentered IV estimator from Theorem 1; RMSE represents the standard deviation of the recentered IV estimator since the bias is negligible. All estimators include an intercept.[30] The columns correspond to the error structures described above, while the rows correspond to the optimal designs, along with a benchmark of a simple Bernoulli experiment with a 50% treatment probability and a recentered IV estimator without whitening. In brackets, we also include the gain in effective sample size, defined as the squared RMSE ratio relative to the benchmark.

All optimal designs beat a standard RCT, which is not optimized to estimate spillover effects. Under homoskedastic errors, the design with $p = \infty$ (i.e., perfectly correlating treatments within blocks) works best. But under more complex and realistic error structures, the advantage of extreme shock correlations deteriorates: the $p = 2$ design is best for the DGP in columns 3–5 with network-dependent or correlated errors. This choice of p also shows robust power gains across all columns, equivalent to increasing sample size of 41–75% in all columns.

To isolate the gain from the optimal design vs. the optimal estimator, Appendix Table A1 reports RMSE of the recentered IV estimator without whitening for the p-optimal designs. On one hand, the $p = \infty$ design involves no whitening yet achieves substantial improvements across most error distributions—suggesting a key role of design. Moreover, for simple (e.g., homoskedastic) errors whitening only reduces precision gains. On the other hand, column 5 shows that with more challenging error distributions extra power gains due to the $p = 1$ and $p = 2$ designs are only realized if the instrument is also whitened. Appendix Table A3 also includes an analysis of coverage

[30] Correspondingly, Algorithm 1 is applied with $R = \mathbf{1}$. We use the optional part of step 2 to block-diagonalize the residualized W. This allows us to solve for the optimal design village by village, with independent randomization across villages.

for confidence intervals built using the normal approximation from Theorem 2. Since the network of Cai et al. (2015) is made up of many independent villages, all designs and simulations have close to nominal coverage, even when shocks are perfectly correlated within villages under $p = \infty$.

4.2 Direct and Indirect Effects: Miguel and Kremer (2004)

Setup. Our second application is in the setting of Miguel and Kremer (2004) who estimate the direct and spillover effects of a randomized school-level deworming treatment on helminth infection rates among children in Kenya. The intervention shocks g_i are at the school level; slightly simplifying the original analysis, we define the outcome y_i as the average infection rate in school i. Thus, the outcome and intervention units are the same set of $N = K = 49$ schools (specifically, Groups 1 and 2 of the experiment for which health outcomes are observed). Following Miguel and Kremer (2004), the spillover treatment x_i is the number (rather than the share) of students in nearby schools that were treated, rescaled by 1,000. This is formalized by the $N \times N$ adjacency matrix W with entry $w_{ik} = E_k/1000$ if school i is within 3km of school $k \neq i$, where E_k is the number of pupils eligible for the treatment in school k, and 0 otherwise (including $w_{ii} = 0$ for all i). The top row of Figure 3(b) shows the exposure matrix W and the matrix of similarity in exposure weights $W'W$.

We calibrate our simulation to a simplified version of column (1) in Table VII of Miguel and Kremer (2004). We estimate a school-level regression $y_i = \alpha + \beta x_i + \tau g_i + \upsilon_i$ using recentered IV without whitening. This yields coefficients $\beta^* = -0.226$ and $\tau^* = -0.258$.[31]

For $S = 1,000$ simulations, we generate the outcome as

$$y_i = \beta^* x_i + \tau^* g_i + \varepsilon_i, \tag{11}$$

where again the distribution of g depends on the experimental design, W is fixed, and we vary $\mathbb{E}_{\mathcal{E}}[\varepsilon\varepsilon']$ according to the same six DGPs described above, with a few minor changes. We replace the village-correlated DGP with a "cluster-correlated" one by creating five clusters of schools using spectral clustering on W and use random cluster shocks. For the network-correlated DGP, we set $W_{\text{raw}} = W$. We calibrate the simulations like in the first application: to match the estimated standard error of the unwhitened recentered IV estimator and targeting the same shares of variation explained by different error components.[32]

Results. We follow Algorithm 1 to solve for the feasible design optimized for the *indirect* effect in a linear model (11) that also includes a direct effect. As in the first application, we include the intercept for all estimators.[33]

[31] Our specification differs from Miguel and Kremer (2004) in several ways. We use a linear model rather than probit, we recenter rather than including additional controls, and we drop a third treatment variable based on exposure to schools between 3 and 6 km away. Despite these simplifications, our estimates are very similar to the corresponding average marginal effects in the original paper, of -0.26 for the indirect effect and -0.25 for the direct effect.

[32] The corresponding parameters are as follows. Homoskedastic: $\nu = 0.17$; heteroskedastic: $\gamma = 0.4$, $\nu = 0.12$; degree-in-mean: $\gamma = 0.056$, $\nu = 0.085$, cluster-correlated: $\nu = 0.10$; network-correlated: $\gamma = 0.12, \nu = 0.05$.

[33] We again use the algorithm with $R = \mathbf{1}$. Here, however, we do not need the optional part of step 2.

Table 2: RMSE and Effective Sample Size Gains for the Miguel and Kremer (2004) Application

(a) Spillover Effect

Design and Estimator:	Homo-skedastic (1)	Hetero-skedastic (2)	Deg.-in-Mean (3)	Cluster-Corr. (4)	Network-Corr. (5)	Estimated Residuals (6)
RCT Benchmark	0.101 [+0%]	0.099 [+0%]	0.128 [+0%]	0.106 [+0%]	0.109 [+0%]	0.179 [+0%]
Optimal, $p = 1$	0.085 [+43%]	0.087 [+30%]	0.087 [+114%]	0.075 [+98%]	0.064 [+191%]	0.120 [+124%]
Optimal, $p = 2$	0.070 [+110%]	0.075 [+76%]	0.069 [+241%]	0.070 [+128%]	0.064 [+190%]	0.114 [+148%]
Optimal, $p = \infty$	0.052 [+277%]	0.056 [+215%]	0.049 [+570%]	0.067 [+147%]	0.057 [+267%]	0.111 [+160%]

(b) Direct Effect

Design and Estimator:	Homo-skedastic (1)	Hetero-skedastic (2)	Deg.-in-Mean (3)	Cluster-Corr. (4)	Network-Corr. (5)	Estimated Residuals (6)
RCT Benchmark	0.052 [+0%]	0.049 [+0%]	0.040 [+0%]	0.041 [+0%]	0.034 [+0%]	0.064 [+0%]
Optimal, $p = 1$	0.054 [-7%]	0.054 [-18%]	0.042 [-7%]	0.046 [-22%]	0.036 [-14%]	0.067 [-10%]
Optimal, $p = 2$	0.054 [-7%]	0.051 [-7%]	0.039 [+6%]	0.045 [-18%]	0.040 [-27%]	0.068 [-12%]
Optimal, $p = \infty$	0.051 [+6%]	0.048 [+5%]	0.040 [+1%]	0.046 [-20%]	0.038 [-23%]	0.100 [-59%]

Notes: For the Miguel and Kremer (2004) application, Panel (a) reports the root mean-squared error of different design-estimator pairs (rows) for the spillover effect under different data-generating processes for the errors (columns). Panel (b) reports the same for the direct effect. The first row is the baseline of independent Bernoulli assignment and unwhitened recentered IV. The other rows are the Algorithm 1 feasible optimal design and optimal IV for different finite values of p (and the analogous solutions for $p = \infty$), with $R = \mathbf{1}$. The text describes the six error DGPs. Increases in effective sample sizes, given in brackets, are computed as the squared ratio of inverse RMSE under the focal scenario relative to the RCT baseline with the same error distributions, minus one.

The bottom row of Figure 3(b) presents the correlation heatmap for the optimal designs used in the simulations. Now that the direct effect is in the model, we no longer have a perfectly correlated design when $p = \infty$. As p decreases, the optimal design from Proposition 1′ weakens both positive and negative correlations of the assignments to protect against adversarially correlated (again, positively or negatively) error distributions.

The RMSE results for the indirect and direct effects are in Table 2, including in brackets the effective sample size gain compared to the baseline RCT design with an unwhitened recentered IV estimator. For estimating the spillover effect, Panel (a) shows that all optimal designs beat the simple RCT under all error structures. Selecting $p = \infty$ gives the best performance in all columns. Error correlations are not adversarial enough for $p = 2$ to dominate the other designs, as it did in the previous application,[34] but it remains a good choice, especially with more complex error structures. Overall, the power gains are even more substantial than in the first application: with the $p = 2$ choice, they range from +76% in column 2 to +241% in column 3, while with $p = \infty$ they reach +570% in column 3.[35] Yet, Appendix Table A4 provides a reason to prefer a conservative choice of p: under $p = \infty$, the resulting covariance matrix of the shocks is too dense to meet the conditions of Theorem 2. For the simulations with correlated errors, confidence intervals built using the normal approximation under the $p = \infty$ design severely undercover the target parameter, while the $p = 1$ design has close to nominal coverage under all DGPs.

Since our proposed designs are optimized for the indirect effect, they can lead to some deterioration of performance of the direct effect estimator, as reported in Panel (b) of Table 2. For $p = \infty$ this deterioration can be substantial, equivalent to a 59% reduction in sample size in column 6. However, for $p = 1$ and $p = 2$ the deterioration is mild, of at most 22% and 27%, respectively.

5 Conclusion

We have shown how researchers can design experiments and choose recentered estimators to more precisely estimate spillover effects, under weak restrictions on the distribution of unobservables. The core insight of this approach is that spillover signal can be increased, by correlating the assignments of connected units, without overly concentrating variation in directions where the unobservables may be adversarially clustered. Balancing this tradeoff between signal and isotropy yields a tractable minimax problem with intuitive special cases and practical approaches to computation. In two semi-synthetic experiments, based on high-profile applications, we find large decreases in spillover effect standard errors relative to conventional randomization and estimation that are possible without large deterioration in the standard errors for the direct effects in the application with such effects.

Several open questions remain. First, while we have focused on linear spillover treatments, and derived theoretical extensions to nonlinear spillover treatment constructions, computation of

[34] A possible reason is that here x_i is the number (rather than share) of treated neighbors, so there is more scope for increasing signal by correlating the assignments—which is what the $p = \infty$ design targets.

[35] To isolate the performance impact of the optimal design vs. the optimal estimator, Appendix Table A2 reports RMSE of the recentered IV estimator without whitening for the p-optimal designs.

optimal designs in nonlinear settings may be challenging. Second, though we have characterized the optimal design and estimator for spillovers in the presence of direct effects, we have not presented the more general problem of minimizing a weighted average of their variances. Third, we have not shown how the optimal design problem changes when a researcher has access to some prior information on the distribution of unobservables, such as from a first-wave pilot (Viviano, 2026). Finally, we have not studied how the minimax approach changes when a researcher is interested in estimating a more complicated parametric model of spillovers, such as from a structural economic model (Borusyak et al., 2025a). We hope to address some of these extensions in future drafts.

References

Armstrong, Timothy B and Michal Kolesár, "Optimal inference in a class of regression models," *Econometrica*, 2018, *86* (2), 655–683.

Atkin, David, Aaron Berman, and Banu Demir, "Reducing Carbon Emissions while Boosting Growth: an Experiment with Turkish Manufacturing Firms," *AEA RCT Registry*, 2024.

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

Baird, Sarah, J Aislinn Bohren, Craig McIntosh, and Berk Özler, "Optimal Design of Experiments in the Presence of Interference," *Review of Economics and Statistics*, 2018, *100* (5), 844–860.

Best, Michael Carlos, Florian Grosset-Touba, Gastón Pierri, Evan Sadler, and Panos Toulis, "Spillover Effects of Tax Enforcement on Production Networks: Evidence from Paraguay. Pre-Analysis Plan," *AEA RCT Registry*, 2025.

Bhatia, Rajendra, *Matrix analysis*, Springer Science & Business Media, 2013.

Bickel, P. J. and Agnes M. Herzberg, "Robustness of Design Against Autocorrelation in Time I: Asymptotic Theory, Optimality for Location and Linear Regression," *The Annals of Statistics*, 1979, *7* (1), 77–95.

Borusyak, Kirill and Peter Hull, "Non-Random Exposure to Exogenous Shocks," *Econometrica*, 2023, *91* (6), 2155–2185.

_ **and** _ , "Negative weights are no concern in design-based specifications," *AEA Papers and Proceedings*, 2024, *114*, 597–600.

_ **and** _ , "Optimal Formula Instruments," *Econometrica*, 2026. Forthcoming.

_ **, Mauricio Cáceres Bravo, and Peter Hull**, "Estimating Demand with Recentered Instruments," *Working Paper*, 2025. arXiv:2504.04056.

_ **, Peter Hull, and Xavier Jaravel**, "Design-Based Identification with Formula Instruments: A Review," *The Econometrics Journal*, 2025, *28* (1), 83–108.

Cai, Jing, Alain de Janvry, and Elisabeth Sadoulet, "Social Networks and the Decision to Insure," *American Economic Journal: Applied Economics*, 2015, *7* (2), 81–108.

Chatterjee, Sourav, "Fluctuations of eigenvalues and second order Poincaré inequalities," *Probability Theory and Related Fields*, 2009, *143* (1), 1–40.

Chaurey, Ritam, Gaurav Nayyar, Siddharth Sharma, and Eric Verhoogen, "Social Learning among Urban Manufacturing Firms: Energy-Efficient Motors in Bangladesh," NBER Working Paper 34296, National Bureau of Economic Research 2025.

Chen, Louis H. Y. and Qi-Man Shao, "Normal approximation under local dependence," *The Annals of Probability*, 2004, *32* (3A), 1985–2028.

Cruces, Guillermo, Dario Tortarolo, and Gonzalo Vazquez-Bare, "Design of partial population experiments with an application to spillovers in tax compliance," *Review of Economics and Statistics*, 2025.

Egger, Dennis, Johannes Haushofer, Edward Miguel, Paul Niehaus, and Michael Walker, "General equilibrium effects of cash transfers: experimental evidence from Kenya," *Econometrica*, 2022, *90* (6), 2603–2643.

Faridani, Stefan and Paul Niehaus, "Linear estimation of global average treatment effects," Technical Report, arXiv 2026. arXiv:2209.14181.

Gao, Mengsi, "Endogenous Interference in Randomized Experiments," 2024.

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

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár, "Contamination Bias in Linear Regressions," *American Economic Review*, 2024, *114* (12), 4015–4051.

Harshaw, Christopher, Fredrik Sävje, David Eisenstat, Vahab Mirrokni, and Jean Pouget-Abadie, "Design and Analysis of Bipartite Experiments under a Linear Exposure-Response Model," *Electronic Journal of Statistics*, 2023, *17* (1), 464–518.

Kallus, Nathan, "On the Optimality of Randomization in Experimental Design: How to Randomize for Minimax Variance and Design-Based Inference," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2021, *83* (2), 404–409.

Kiefer, J, "General Equivalence Theory for Optimum Designs (Approximate Theory)," *The Annals of Statistics*, 1974, *2* (5), 849–879.

Kiefer, Jack and Jacob Wolfowitz, "Optimum designs in regression problems," *The Annals of Mathematical Statistics*, 1959, *30* (2), 271–294.

Lei, Lihua, Peter J Bickel, and Noureddine El Karoui, "Asymptotics for high dimensional regression M-estimates: fixed design results," *Probability Theory and Related Fields*, 2018, *172* (3), 983–1079.

Lewis, Adrian S, "The convex analysis of unitarily invariant matrix functions," *Journal of Convex Analysis*, 1995, *2* (1), 173–183.

—, "Derivatives of spectral functions," *Mathematics of Operations Research*, 1996, *21* (3), 576–588.

Miguel, Edward and Michael Kremer, "Worms: Identifying impacts on education and health in the presence of treatment externalities," *Econometrica*, 2004, *72* (1), 159–217.

Pouget-Abadie, Jean, Kevin Aydin, Warren Schudy, Kay Brodersen, and Vahab Mirrokni, "Variance Reduction in Bipartite Experiments through Correlation Clustering," in "Advances in Neural Information Processing Systems," Vol. 32 2019.

Sun, Liyang, "Empirical welfare maximization with constraints," *Journal of Econometrics*, 2026, *253*, 106169.

Thiyageswaran, Vydhourie, Alex Kokot, Jennifer Brennan, Marina Meila, Christina Lee Yu, and Maryam Fazel, "Optimal Design under Interference, Homophily, and Robustness Trade-offs," *arXiv preprint*, 2026. arXiv:2601.17145 [stat].

Viviano, Davide, "Experimental Design under Network Interference," *Journal of Political Economy Microeconomics*, 2026. Forthcoming.

_ , Lihua Lei, Guido Imbens, Brian Karrer, Okke Schrijvers, and Liang Shi, "Causal clustering: design of cluster experiments under network interference," *Working Paper*, 2026. arXiv:2310.14983.

Zenou, Yves, "Peer vs. Network Effects: Microfoundations, Identification, and Beyond," IZA Discussion Paper 18501, IZA Institute of Labor Economics 2026.

Appendix

A Proofs of Main Results

A.1 Theorem 1

We drop the δ subscript from $\tilde{x}$, S, and z^* to simplify notation. Write $\Omega_\varepsilon = \mathbb{E}_{\mathcal{E}}[\varepsilon\varepsilon']$ and, for any recentered instrument z, $Q_z = \mathbb{E}_\delta[zz']$. The approximate variance of the recentered IV estimator can be written

$$\mathcal{V}_{\delta,\mathcal{E}}[z] = \frac{\mathbb{E}_\delta[z'\varepsilon\varepsilon' z]}{\mathbb{E}_\delta[z'x]^2} = \frac{\operatorname{tr}(Q_z\Omega_\varepsilon)}{\mathbb{E}_\delta[z'\tilde{x}]^2}$$

where the denominator uses $\mathbb{E}_\delta[z'\mathbb{E}_\delta[x]] = 0$. By the duality of Schatten p- and q-norms where $q \in [1,\infty]$, such that $1/p + 1/q = 1$, the minimax approximate variance satisfies:

$$\max_{\mathcal{E}\in\mathcal{F}_p(\sigma)} \mathcal{V}_{\delta,\mathcal{E}}[z] = \sup_{\Omega_\varepsilon \succeq 0, \|\Omega_\varepsilon\|_p \le N^{1/p}\sigma^2} \frac{\operatorname{tr}(Q_z\Omega_\varepsilon)}{\mathbb{E}_\delta[z'\tilde{x}]^2} = N^{1/p}\sigma^2 \frac{\|Q_z\|_q}{\mathbb{E}_\delta[z'\tilde{x}]^2}. \tag{A1}$$

Note that the additional constraint that Ω_ε is positive semi-definite does not affect the supremum because $Q_z \succeq 0$.

Next, we propose a worst-case Ω_ε for a given z and show that it attains the supremum. Fix an instrument z for which the second moment is non-zero, so that $\|Q_z\|_q \neq 0$. Then, conjecture the following worst-case matrix :

$$\Omega_\varepsilon^* = \begin{cases} N\sigma^2 \frac{B_z}{\operatorname{tr}(B_z)}, & p = 1, \\ N^{1/p}\sigma^2 \frac{Q_z^{q-1}}{\|Q_z\|_q^{q-1}}, & 1 < p < \infty, \\ \sigma^2 Q_z^0, & p = \infty, \end{cases} \tag{A2}$$

where $B_z \neq 0$ is any positive semidefinite matrix with $\operatorname{Im}(B_z) \subseteq \mathcal{M}_z$ where $\mathcal{M}_z$ is the eigenspace associated with the largest eigenvalue of Q_z. For $1 < p < \infty$, $p(q-1) = q$ and $q/p = q-1$, so:

$$\begin{aligned} \|Q_z^{q-1}\|_p &= \left\{\operatorname{tr}(Q_z^{p(q-1)})\right\}^{1/p} \\ &= \{\operatorname{tr}(Q_z^q)\}^{1/p} \\ &= \|Q_z\|_q^{q-1}. \end{aligned}$$

This means $\|\Omega_\varepsilon^*\|_p = N^{1/p}\sigma^2$ so $\Omega_\varepsilon^* \in \mathcal{F}_p(\sigma)$ is feasible. In addition,

$$\begin{aligned}\operatorname{tr}(Q_z\Omega_\varepsilon^*) &= N^{1/p}\sigma^2\frac{\operatorname{tr}(Q_z^q)}{\|Q_z\|_q^{q-1}} \\ &= N^{1/p}\sigma^2\frac{\|Q_z\|_q^q}{\|Q_z\|_q^{q-1}} \\ &= N^{1/p}\sigma^2\|Q_z\|_q,\end{aligned}$$

which attains the supremum from the optimization problem in (A1). Similarly, for $p=1$ and $q=\infty$, we have $\|\Omega_\varepsilon^*\|_1 = N\sigma^2$ and the matrix is feasible. Furthermore, we have $Q_zB_z = \|Q_z\|_\infty B_z$, so $\operatorname{tr}(Q_z\Omega_\varepsilon^*) = N\sigma^2\|Q_z\|_\infty$, which attains the supremum. For $p=\infty$ and $q=1$, $\|\Omega_\varepsilon^*\|_\infty = \sigma^2$ and so it is again feasible, and $\operatorname{tr}(Q_z\Omega_\varepsilon^*) = \sigma^2\operatorname{tr}(Q_z) = \sigma^2\|Q_z\|_1$ which again attains the supremum.

To find the optimal instrument, we proceed in two steps. We first show that there is an instrument that minimizes (A1) that takes the form $z = V\tilde{x}$. We then find the optimal V matrix. Define the linear projection of z on $\tilde{x}$ as $v_z = V_z\tilde{x}$ for $V_z = \mathbb{E}_\delta[z\tilde{x}']S^\dagger$, with the residual $u_z = z - v_z$ that satisfies $\mathbb{E}_\delta[\tilde{x}u_z'] = \mathbb{E}_\delta[v_zu_z'] = 0$. Thus,

$$\mathbb{E}_\delta\left[v_z\tilde{x}'\right] = \mathbb{E}_\delta\left[(z-u_z)\,\tilde{x}'\right] = \mathbb{E}_\delta\left[\widetilde{zx'}\right]$$

and

$$Q_z = \mathbb{E}_\delta\left[(v_z+u_z)(v_z+u_z)'\right] = \mathbb{E}_\delta\left[v_zv_z'\right] + \mathbb{E}_\delta\left[u_zu_z'\right] \succeq \mathbb{E}_\delta\left[v_zv_z'\right] = Q_v.$$

By the monotonicity of the Schatten q-norm on the positive semidefinite cone, $\|Q_z\|_q \geq \|Q_v\|_q$.

Thus, the objective in (A1) is weakly lower for the instrument v than z and it is without loss to restrict attention to the instruments that take the form $z = V\tilde{x}$. For them, $Q_z = VSV'$ and $\mathbb{E}_\delta[z'\tilde{x}] = \operatorname{tr}(VS)$ such that the optimal V matrix solves

$$\min_V \frac{\|VSV'\|_q}{\operatorname{tr}(VS)^2}. \tag{A3}$$

Let $H = VS^{1/2}$ and $r = 2p/(p+1) \in [1,2]$ be the Hölder conjugate of $2q$, so that $\frac{1}{r} + \frac{1}{2q} = 1$. By the definition of the Schatten norm, $\|VSV'\|_q = \|HH'\|_q = \|H\|_{2q}^2$ while $\operatorname{tr}(VS) = \operatorname{tr}\left(HS^{1/2}\right)$. By the Hölder inequality for Schatten norms (Corollary IV.2.6 of Bhatia (2013)),

$$\operatorname{tr}(VS)^2 \leq \|H\|_{2q}^2\left\|S^{1/2}\right\|_r^2 = \|VSV'\|_q\cdot\left\{\operatorname{tr}\left(S^{r/2}\right)\right\}^{2/r}. \tag{A4}$$

Plugging this inequality into (A3), we have that for any V, $\frac{\|VSV'\|_q}{\operatorname{tr}(VS)^2} \geq \left\{\operatorname{tr}\left(S^{r/2}\right)\right\}^{-2/r}$. This lower bound is attained by $V^* = \left(S^\dagger\right)^{1-r/2} = \left(S^\dagger\right)^{1/(p+1)}$ since $\|V^*SV^{*\prime}\|_q = \operatorname{tr}\left(S^{p/(p+1)}\right)^{(p-1)/p}$ for $p>1$ and $\|V^*SV^{*\prime}\|_q = 1$ for $p=1$, while $\operatorname{tr}(V^*S) = \operatorname{tr}\left(S^{p/(p+1)}\right)$. Therefore, V^* solves (A3). The optimal instrument is then $z^* = \left(S^\dagger\right)^{1/(p+1)}\tilde{x}$ and by plugging in the optimum of (A3) into (A1), we have $\max_{\mathcal{E}\in\mathcal{F}_p(\sigma)}\mathcal{V}_{\delta,\mathcal{E}}[z^*] = N^{1/p}\sigma^2\left\{\operatorname{tr}\left(S^{r/2}\right)\right\}^{-2/r}$.

To get an optimal matrix Ω_ε^* corresponding to z^*, we use (A2). For $1 < p < \infty$, we have

$$\Omega_\varepsilon^* = N^{1/p}\sigma^2 \frac{Q_{z^*}^{q-1}}{\|Q_{z^*}\|_q^{q-1}}.$$

Then, plugging in the expression for $Q_{z^*} = S^{(p-1)/(p+1)}$, we have that the worst-case error covariance matrix corresponding to z^* is

$$\Omega_\varepsilon^* = N^{1/p}\sigma^2 \frac{S^{1/(p+1)}}{\left(\operatorname{tr}\left(S^{p/(p+1)}\right)\right)^{1/p}}.$$

For $p = 1$, $Q_{z^*} = S^0$. The eigenspace associated with the largest eigenvalue of Q_z is $\operatorname{Im}(S)$, so to match the formula for general p while still meeting the conditions of (A2) we can choose $B_z = S^{1/2}$ and $\Omega_\varepsilon^* = N\sigma^2 \frac{S^{1/2}}{\operatorname{tr}(S^{1/2})}$. The worst-case matrix is not unique here, because B_z is not unique ($B_z = S$ also works, for example), but it is convenient to match the general formula. Finally, for $p = \infty$, $\Omega_\varepsilon^* = \sigma^2 S^0$.

A.2 Proposition 1

Follows immediately from the expression for the worst-case approximate variance in Theorem 1, which is finite whenever $S_\delta \neq 0$.

A.3 Proposition 2

For the first claim, take any δ and define its complement δ^c by drawing $\mathbf{1} - g$ whenever δ would draw g. Now symmetrize:

$$\delta^{sym} = \frac{1}{2}\delta + \frac{1}{2}\delta^c.$$

Clearly $\mathbb{E}_{\delta^{sym}}[g_k] = 0.5$ for all k while

$$\operatorname{Var}_{\delta^{sym}}[g] = \operatorname{Var}_\delta[g] + \left(\mathbb{E}_\delta[g] - \frac{1}{2}\mathbf{1}\right)\left(\mathbb{E}_\delta[g] - \frac{1}{2}\mathbf{1}\right)' \succeq \operatorname{Var}_\delta[g].$$

Hence

$$S_{\delta^{sym}} = W\operatorname{Var}_{\delta^{sym}}[g]\,W' \succeq W\operatorname{Var}_\delta[g]\,W' = S_\delta.$$

Furthermore, since $A \mapsto \operatorname{tr}(A^{p/(p+1)})$ is increasing on the positive semidefinite cone,

$$\operatorname{tr}\left(S_{\delta^{sym}}^{p/(p+1)}\right) \geq \operatorname{tr}\left(S_\delta^{p/(p+1)}\right).$$

Hence symmetrizing weakly improves the Proposition 1 objective, so some δ^* has $\mathbb{E}_{\delta^*}[g_k] = 0.5$ for all k.

For the second claim, suppose $W = \text{diag}(W_{(1)}, \ldots, W_{(B)})$ for blocks $W_{(b)} \in \mathbb{R}^{N_b \times K_b}$. For any design, let δ_b be the marginal distribution of $g_{(b)}$. Define the product design $\delta^{ind} = \delta_1 \otimes \cdots \otimes \delta_B$ and note

$$S_{\delta^{ind}} = \text{diag}(S_{\delta,11}, \ldots, S_{\delta,BB})$$

for $S_{\delta,bb} = W_{(b)} \text{Var}_\delta \left[g_{(b)}\right] W'_{(b)}$ since cross-block covariances vanish under the product design. $S_{\delta^{ind}}$ is a pinching of S_δ, following the definition in Chapter 2 of Bhatia (2013). By Equation II.39 of Bhatia (2013), we have a majorization relationship $\lambda(S_{\delta^{ind}}) \prec \lambda(S_\delta)$, where $\lambda(A)$ gives the vector of eigenvalues of matrix A. Since $x \mapsto x^{p/(p+1)}$ is concave for $p \in [1, \infty]$, by the characterization of majorization in Theorem II.3.1 of Bhatia (2013), we have:

$$\text{tr}(S_{\delta^{ind}}^{p/(p+1)}) \geq \text{tr}(S_\delta^{p/(p+1)}).$$

Hence any design is weakly improved by replacing it with the product of its block marginals and we may restrict attention to product designs.

For the third claim, we have

$$\text{tr}(S_{\delta^{ind}}^{p/(p+1)}) = \sum_{b=1}^{B} \text{tr}(S_{\delta^{ind},bb}^{p/(p+1)}).$$

Thus, the design can be optimized block by block.

A.4 Lemma 1, Theorem 1′, Propositions 1′–2′

Lemma 1. Take any $\boldsymbol{z} \in \mathcal{Z}_\delta^{(2)}$. If $M_{\boldsymbol{z}} = \mathbb{E}_\delta\left[\boldsymbol{x}'\boldsymbol{z}\right]$ is nonsingular, we can define $z = \boldsymbol{z} M_{\boldsymbol{z}}^{-1} e_1$, where $e_1 = (1, 0)'$. Since $\mathbb{E}_\delta[\boldsymbol{z}] = 0$, $\mathbb{E}_\delta[z] = 0$, so it is re-centered. Also, $\mathbb{E}_\delta\left[\boldsymbol{x}'z\right] = M_z M_z^{-1} e_1 = e_1$, i.e. $\mathbb{E}_\delta\left[x'z\right] = 1$ and $\mathbb{E}_\delta\left[h'z\right] = 0$. Next, to show the variance equality, we have that $\varepsilon' z = \varepsilon' \boldsymbol{z} M_z^{-1} e_1$. Because $\mathbb{E}_\delta\left[x'z\right] = 1$,

$$\begin{aligned}
\mathcal{V}_{\delta,\mathcal{E}}\left[z\right] &= \text{Var}_{\delta,\mathcal{E}}\left[\varepsilon' \boldsymbol{z} M_z^{-1} e_1\right] \\
&= e_1' \left(M_{\boldsymbol{z}}^{-1}\right)' \text{Var}_{\delta,\mathcal{E}}\left[\boldsymbol{z}'\varepsilon\right] M_{\boldsymbol{z}}^{-1} e_1 \\
&= \boldsymbol{\mathcal{V}}_{\delta,\mathcal{E}}\left[\boldsymbol{z}\right].
\end{aligned}$$

Hence every vector instrument has a scalar instrument orthogonal to h with the same approximate variance for the spillover effect. If M_z is singular, $\boldsymbol{\mathcal{V}}_{\delta,\mathcal{E}}\left[\boldsymbol{z}\right] = \infty$ so we can take $z = 0$ which achieves the same approximate variance.

Conversely, take any $z \in \mathcal{Z}_\delta$ with $\mathbb{E}_\delta\left[z'h\right] = 0$. Let $\boldsymbol{z} = (z, h - \mathbb{E}_\delta\left[h\right])$ and note that $\mathbb{E}_\delta\left[\boldsymbol{z}\right] = 0$, so $\boldsymbol{z} \in \mathcal{Z}_\delta^{(2)}$. If

$$\mathbb{E}_\delta\left[\boldsymbol{z}'\boldsymbol{x}\right] = \begin{bmatrix} \mathbb{E}_\delta\left[z'x\right] & 0 \\ \mathbb{E}_\delta\left[(h - \mathbb{E}_\delta\left[h\right])'x\right] & \mathrm{tr}(\mathrm{Var}_\delta\left[h\right]) \end{bmatrix}$$

is nonsingular, simple algebra verifies that

$$\mathcal{V}_{\delta,\mathcal{E}}\left[z\right] = \boldsymbol{\mathcal{V}}_{\delta,\mathcal{E}}\left[\boldsymbol{z}\right].$$

If $\mathbb{E}_\delta\left[\boldsymbol{z}'\boldsymbol{x}\right]$ is singular, from $\mathrm{tr}(\mathrm{Var}_\delta\left[h\right]) > 0$ we conclude that $\mathbb{E}_\delta\left[z'x\right] = 0$. Thus, we have $\mathcal{V}_{\delta,\mathcal{E}}\left[z\right] = \boldsymbol{\mathcal{V}}_{\delta,\mathcal{E}}\left[\boldsymbol{z}\right] = \infty$.

Theorem 1′. By Lemma 1, it is enough to solve

$$\inf_{z\in\mathcal{Z}_\delta:\,\mathbb{E}_\delta[z'h]=0}\ \max_{\mathcal{E}\in\mathcal{F}_p(\sigma)} \mathcal{V}_{\delta,\mathcal{E}}\left[z\right]. \tag{A5}$$

By (A1), this problem can be simplified to

$$\inf_{z\in\mathcal{Z}_\delta:\,\mathbb{E}_\delta[z'h]=0} N^{1/p}\sigma^2\frac{\left\|Q_z\right\|_q}{\mathbb{E}_\delta\left[z'\tilde{x}\right]^2}. \tag{A6}$$

Since $\mathbb{E}_\delta\left[z\right] = 0$, for $\tilde{h} = h - \mathbb{E}_\delta\left[h\right]$, $\mathbb{E}_\delta\left[z'h\right] = \mathbb{E}_\delta\left[z'\tilde{h}\right]$, and similarly, $\mathbb{E}_\delta\left[z'x\right] = \mathbb{E}_\delta\left[z'\tilde{x}\right]$. This problem is invariant to rescaling of z: for any scalar $\alpha \neq 0$, $Q_{\alpha z} = \alpha^2 Q_z$ and $\mathbb{E}_\delta\left[(\alpha z)'\,\tilde{x}\right]^2 = \alpha^2\mathbb{E}_\delta\left[z'\tilde{x}\right]^2$. Thus, it is without loss to impose the constraint $\left\|Q_z\right\|_q \leq 1$ and replace $\left\|Q_z\right\|_q$ with its maximum value of one in (A6): i.e.,

$$\inf_{z\in\mathcal{Z}_\delta:\,\mathbb{E}_\delta[z'h]=0} N^{1/p}\sigma^2\frac{\left\|Q_z\right\|_q}{\mathbb{E}_\delta\left[z'\tilde{x}\right]^2} = N^{1/p}\sigma^2\Big(\max_{z\in\mathcal{Z}_\delta:\mathbb{E}_\delta\left[z'\tilde{h}\right]=0,\ \left\|Q_z\right\|_q\leq 1} \mathbb{E}_\delta\left[z'\tilde{x}\right]\Big)^{-2}.$$

We next drop constants and the exponent, which do not affect the solution, and convert the constraint $\mathbb{E}_\delta\left[z'\tilde{h}\right] = 0$ into the Lagrangian form:

$$\max_{z\in\mathcal{Z}_\delta:\mathbb{E}_\delta\left[z'\tilde{h}\right]=0,\ \left\|Q_z\right\|_q\leq 1} \mathbb{E}_\delta\left[z'\tilde{x}\right] = \max_{z\in\mathcal{Z}_\delta:\left\|Q_z\right\|_q\leq 1}\ \inf_{c\in\mathbb{R}} \mathbb{E}_\delta\left[z'(\tilde{x} - c\tilde{h})\right].$$

We now apply the Sion minimax theorem to swap the order of optimization, noting that the set $\mathcal{C} = \left\{z \in \mathcal{Z}_\delta : \left\|Q_z\right\|_q \leq 1\right\}$ is convex (since $\mathcal{Z}_\delta$ is convex and by the properties of the Schatten norm) and compact (because the support of g is finite), that $\mathbb{R}$ is convex, that $z'(\tilde{x} - c\tilde{h})$ is linear in z and affine in c, and that the expectation is a linear operator:

$$\max_{z\in\mathcal{Z}_\delta:\left\|Q_z\right\|_q\leq 1}\ \inf_{c\in\mathbb{R}} \mathbb{E}_\delta\left[z'(\tilde{x} - c\tilde{h})\right] = \inf_{c\in\mathbb{R}}\ \max_{z\in\mathcal{Z}_\delta:\left\|Q_z\right\|_q\leq 1} \mathbb{E}_\delta\left[z'(\tilde{x} - c\tilde{h})\right]. \tag{A7}$$

Although their optimal values are equal, we have not yet shown that solutions for one side are optimal for the other side. We proceed by finding a solution to the right side problem, and showing that it is in fact optimal for the left-side problem. Applying the scale-invariance argument again, this problem is equivalent to

$$\inf_{c\in\mathbb{R}}\left(\min_{z\in\mathcal{Z}_\delta} N^{1/p}\sigma^2 \frac{\|Q_z\|_q}{\mathbb{E}_\delta\left[z'(\tilde{x}-c\tilde{h})\right]^2}\right)^{-1/2}. \tag{A8}$$

For fixed c, since $\tilde{x}-c\tilde{h}=(W-cU)\tilde{g}$, the inner optimization problem is therefore the Theorem 1 problem applied to the residualized exposure matrix $W-cU$. So we can write:

$$\min_{z\in\mathcal{Z}_\delta} N^{1/p}\sigma^2 \frac{\|Q_z\|_q}{\mathbb{E}_\delta\left[z'(\tilde{x}-c\tilde{h})\right]^2} = N^{1/p}\sigma^2\left\{\operatorname{tr}\left(\left[(W-cU)\Sigma_\delta(W-cU)'\right]^{p/(p+1)}\right)\right\}^{-(p+1)/p}. \tag{A9}$$

We next show that the outer optimization in (A8), which can be equivalently written as $\inf_{c\in\mathbb{R}} H(c)$ for $H(c)=\operatorname{tr}\Big([(W-cU)\Sigma_\delta(W-cU)']^{p/(p+1)}\Big)$, attains its minimum, so that c_δ and $S_\delta^\perp$ are well-defined. This follows because we have assumed that $\operatorname{tr}(U\Sigma_\delta U')>0$, and thus $H(c)\to\infty$ when $|c|\to\infty$, while $H(c)$ is continuous in c as a composition of continuous functions. For a fixed $c=c_\delta$, we know from Theorem 1 with W replaced by $W-c_\delta U$, that the minimal approximate variance is $\left\{\operatorname{tr}\left([(W-c_\delta U)\Sigma_\delta(W-c_\delta U)']^{p/(p+1)}\right)\right\}^{-(p+1)/p}=\left\{\operatorname{tr}\left[(S_\delta^\perp)^{p/(p+1)}\right]\right\}^{-(p+1)/p}$, which completes the proof of the first part of the Theorem. For the optimal IV, an optimal recentered instrument from Theorem 1 is

$$z_\delta^*(g)=\left((S_\delta^\perp)^\dagger\right)^{1/(p+1)}\tilde{x}_\delta^\perp.$$

This $z_\delta^*(g)$ is a solution to the right-side problem in (A7). For it to also be a solution to the left-side problem in (A7) and therefore to the original problem, it is enough to show that it is feasible in the original problem. It is clearly recentered, so it remains to verify that the proposed solution $z_\delta^*(g)$ is orthogonal to h. Let $z_c=\left((S_{\delta,c})^\dagger\right)^{1/(p+1)}(\tilde{x}-c\tilde{h})$, where $S_{\delta,c}=(W-cU)\Sigma_\delta(W-cU)'$. Let $d=\operatorname{rank}(\Sigma_\delta)$ and choose a factor $R_\delta\in\mathbb{R}^{K\times d}$ satisfying $R_\delta R_\delta'=\Sigma_\delta$. Define$A_c=(W-cU)R_\delta$, so $S_c=A_cA_c'$ and

$$H(c)=\operatorname{tr}\left(S_c^{p/(p+1)}\right)=\|A_c\|_r^r$$

for $r=2p/(p+1)$ as in Appendix A.1. Let $s\in\mathbb{R}^m$, where $m=\min\{N,d\}$, and define $f_r(s)=\sum_{j=1}^m |s_j|^r$. $H(c)$ is the function $f_r(\cdot)$ applied to the singular values of A_c. Since $f_r(\cdot)$ is invariant to sign changes and permutations, by Theorem 3.1 of Lewis (1995), if $f_r(s)$ is differentiable at the vector of singular values,

$$\nabla_A\|A_c\|_r^r = r\left((A_cA_c')^\dagger\right)^{1-r/2}A_c,$$

When $p>1$ and therefore $r>1$, $f_r(s)$ is always differentiable. But when $p=1$, $f_1(s)$ is differentiable only if all singular values of A_{c_δ} are non-zero. A sufficient condition is the rank condition that we have assumed: $\text{rank}((W - c_\delta U)R_\delta) = \text{rank}\big((W - c_\delta U)\Sigma^{1/2}\big) = \min\{N, \text{rank}(\Sigma_\delta)\}$. By the chain rule, given the derivative of the matrix function above and that $dA_c/dc = -UR_\delta$, we have shown that $H(c)$ is always differentiable at c_δ.

Since z_c, up to a normalization, is a solution to the inner maximization of the right side of (A7), the envelope theorem implies $H'(c) \propto \mathbb{E}_\delta \left[z_c'\tilde{h}\right]$. Since $H'(c_\delta) = 0$, we have that $\mathbb{E}\left[z_\delta^*(g)'\tilde{h}\right] = 0$, as required for feasibility of $z_\delta^*(g)$ in the original problem.

Finally, Lemma 1 maps this constrained scalar optimum back to the two-instrument problem: it is implemented by $\boldsymbol{z}_i^*(g) = (z_\delta^*(g)_i, h_i - \mathbb{E}_\delta[h_i])$, with the same optimal value, where $\text{tr}(U\Sigma_\delta U') > 0$ guarantees that the first-stage matrix is non-singular.

Proposition 1′. Follows immediately from Theorem 1′.

Proposition 2′. Follows analogously to Proposition 2.

A.5 Characterization of Relaxed Solution from Section 3.1

We first note that, since W has no zero columns, $W'W$ has all positive diagonal elements. Thus, constraining $\ell^* > 0$ is without loss since $\text{tr}\left(M_\ell^p\right) \to \infty$ whenever any ℓ_k approaches zero. Next, we establish:

Lemma A1. *ℓ^* solves (8) if and only if it satisfies:*

$$\ell_k^* = \frac{\left(M_{\ell^*}^p\right)_{kk}}{\text{tr}\left(M_{\ell^*}^p\right)}, \qquad k = 1, \dots, K.$$

Proof. We show below that the objective function in (8) is convex. Furthermore, Slater's condition holds since the point $\bar{\ell} = 1/K \cdot \mathbf{1}_K$ is strictly feasible. Thus, the KKT condition is necessary and sufficient. Some $\ell \in \Delta_K$ minimizes $f(\ell) = \text{tr}\left(M_\ell^p\right)$ subject to $\sum_k \ell_k = 1$ if and only if

$$\frac{\partial\, \text{tr}\left(M_\ell^p\right)}{\partial \ell_k} = \text{const}, \qquad k = 1, \dots, K. \tag{A10}$$

By the chain rule and the derivative of spectral functions (i.e., functions of symmetric matrices that depend only on the eigenvalues of the matrix) given in Theorem 1.1 of Lewis (1996),

$$\frac{\partial}{\partial \ell_k} \text{tr}(M_\ell^p) = p\ \text{tr}\left(M_\ell^{p-1}\frac{\partial M_\ell}{\partial \ell_k}\right).$$

Since $\frac{\partial M_\ell}{\partial \ell_k} = -\frac{1}{2\ell_k}\left(e_k e_k' M_\ell + M_\ell e_k e_k'\right)$, we have $\frac{\partial}{\partial \ell_k}\text{tr}(M_\ell^p) = -p\frac{(M_\ell^p)_{kk}}{\ell_k}$. By (A10), $\ell_k^* \propto \left(M_{\ell^*}^p\right)_{kk}$. Using $\sum_k \ell_k^* = 1$ and $\sum_k \left(M_{\ell^*}^p\right)_{kk} = \text{tr}\left(M_{\ell^*}^p\right)$ yields the desired characterization.

To verify convexity, let $t \in [0,1]$ and $G_\ell = WD_\ell^{-1}W'$. Since the matrices G_ℓ and M_ℓ have the same nonzero eigenvalues, $f(\ell) = \mathrm{tr}(G_\ell^p)$. Writing $w_{\cdot k}$ for the k-th column of W, we have

$$G_\ell = \sum_{k=1}^{K} \frac{w_{\cdot k} w_{\cdot k}'}{\ell_k}.$$

For any $\ell, \tilde{\ell} > 0$, the convexity of $x \mapsto x^{-1}$ gives $G_{t\ell+(1-t)\tilde{\ell}} \preceq tG_\ell + (1-t)G_{\tilde{\ell}}$. The function $\mathrm{tr}(A^p) = \|A\|_p^p$ is convex and monotonic on the positive semidefinite cone for $p \geq 1$, since it is the composition of a norm with a convex function. Applying these arguments,

$$\begin{aligned} f(t\ell + (1-t)\tilde{\ell}) &= \|G_{t\ell+(1-t)\tilde{\ell}}\|_p^p \\ &\leq \|tG_\ell + (1-t)G_{\tilde{\ell}}\|_p^p \\ &\leq t\|G_\ell\|_p^p + (1-t)\|G_{\tilde{\ell}}\|_p^p \\ &= tf(\ell) + (1-t)f(\tilde{\ell}). \end{aligned}$$

Thus, the objective is convex. □

Proof of Proposition 3. Define

$$\bar{W}_\ell = WD_\ell^{-1/2}, \qquad \bar{\Sigma}_\ell = D_\ell^{1/2}\Sigma D_\ell^{1/2},$$

such that

$$W\Sigma W' = \bar{W}_\ell \bar{\Sigma}_\ell \bar{W}_\ell'.$$

If $\Sigma \in \mathcal{Q}_K$, then

$$\mathrm{tr}(\bar{\Sigma}_\ell) = \sum_{k=1}^{K} \ell_k \Sigma_{kk} = \frac{1}{4}.$$

Using the Hölder inequality for Schatten norms (Corollary IV.2.6 of Bhatia (2013)) with $\left(\frac{2p}{p+1}\right)^{-1} = (2p)^{-1} + 2^{-1}$, we have

$$\mathrm{tr}\left((W\Sigma W')^{p/(p+1)}\right) = \left\|\bar{W}_\ell \bar{\Sigma}_\ell^{1/2}\right\|_{2p/(p+1)}^{2p/(p+1)} \leq \left\|\bar{W}_\ell\right\|_{2p}^{2p/(p+1)} \cdot \left\|\bar{\Sigma}_\ell^{1/2}\right\|_2^{2p/(p+1)}.$$

Since

$$\left\|\bar{\Sigma}_\ell^{1/2}\right\|_2^2 = \mathrm{tr}(\bar{\Sigma}_\ell) = \frac{1}{4}$$

and

$$\left\|\bar{W}_\ell\right\|_{2p}^{2p} = \mathrm{tr}\left((\bar{W}_\ell' \bar{W}_\ell)^p\right) = \mathrm{tr}(M_\ell^p),$$

it follows that

$$\mathrm{tr}\left((W\Sigma W')^{p/(p+1)}\right) \leq 4^{-p/(p+1)}\, \mathrm{tr}(M_\ell^p)^{1/(p+1)}.$$

This bound holds for every $\ell \in \operatorname{int}(\Delta_K)$, so

$$\operatorname{tr}\left((W\Sigma W')^{p/(p+1)}\right) \leq 4^{-p/(p+1)} \operatorname{tr}(M_{\ell^*}^p)^{1/(p+1)}. \tag{A11}$$

It remains to show that Σ^* belongs to $\mathcal{Q}_K$ and attains the upper bound in (A11). We have $\Sigma^* \succeq 0$ because $M_{\ell^*}^p \succeq 0$. Moreover, using Lemma A1,

$$\Sigma_{kk}^* = \frac{1}{4}\frac{\left(M_{\ell^*}^p\right)_{kk}}{\ell_k^* \operatorname{tr}\left(M_{\ell^*}^p\right)} = \frac{1}{4}.$$

Plugging in Σ^* into the objective yields

$$\operatorname{tr}\left((W\Sigma^* W')^{p/(p+1)}\right) = \operatorname{tr}\left((\bar{W}_{\ell^*}\bar{\Sigma}_{\ell^*}^*\bar{W}_{\ell^*}')^{p/(p+1)}\right)$$

where $\bar{\Sigma}_{\ell^*}^* = \frac{1}{4}\frac{M_{\ell^*}^p}{\operatorname{tr}\left(M_{\ell^*}^p\right)}$. The non-zero eigenvalues of $\bar{W}_{\ell^*}\bar{\Sigma}_{\ell^*}^*\bar{W}_{\ell^*}'$ equal the non-zero eigenvalues of $\left(\bar{\Sigma}_{\ell^*}^*\right)^{1/2}\bar{W}_{\ell^*}'\bar{W}_{\ell^*}\left(\bar{\Sigma}_{\ell^*}^*\right)^{1/2}$. Noting that $\bar{W}_{\ell^*}'\bar{W}_{\ell^*} = M_{\ell^*}$, the latter matrix equals $\frac{M_{\ell^*}^{p/2}}{2\sqrt{\operatorname{tr}\left(M_{\ell^*}^p\right)}} M_{\ell^*} \frac{M_{\ell^*}^{p/2}}{2\sqrt{\operatorname{tr}\left(M_{\ell^*}^p\right)}} = \frac{M_{\ell^*}^{p+1}}{4\operatorname{tr}\left(M_{\ell^*}^p\right)}$. Thus, non-zero eigenvalues of $\left(\bar{W}_{\ell^*}\bar{\Sigma}_{\ell^*}^*\bar{W}_{\ell^*}'\right)^{p/(p+1)}$ equal those of $\left(\frac{M_{\ell^*}^{p+1}}{4\operatorname{tr}\left(M_{\ell^*}^p\right)}\right)^{p/(p+1)} = \frac{M_{\ell^*}^p}{\left(4\operatorname{tr}\left(M_{\ell^*}^p\right)\right)^{p/(p+1)}}$, from which we conclude that

$$\operatorname{tr}\left((W\Sigma^* W')^{p/(p+1)}\right) = \operatorname{tr}\left(\frac{M_{\ell^*}^p}{\left(4\operatorname{tr}\left(M_{\ell^*}^p\right)\right)^{p/(p+1)}}\right) = 4^{-p/(p+1)} \operatorname{tr}(M_{\ell^*}^p)^{1/(p+1)}, \tag{A12}$$

achieving the upper bound in (A11). This concludes the proof.

Proof of Corollary 1 When $p = 1$,

$$\operatorname{tr}(M_\ell^p) = \operatorname{tr}(M_\ell) = \sum_{k=1}^K \frac{\|w_{\cdot k}\|_2^2}{\ell_k}.$$

Thus, the weights solving (8) satisfy

$$\ell_k^* \propto \|w_{\cdot k}\|_2.$$

Plugging these weights into the formula for Σ^* yields the corollary result.

Proof of Corollary 2 We show that, under the conditions of the corollary, $\ell^* = (1/K)\cdot \mathbf{1}_K$. This result follows by Lemma A1: with such ℓ^*, $D_{\ell^*} = \frac{1}{K}I_K$, $M_{\ell^*}^p = (KW'W)^p$, and thus

$$\frac{\left(M_{\ell^*}^p\right)_{kk}}{\operatorname{tr}\left(M_{\ell^*}^p\right)} = \frac{((W'W)^p)_{kk}}{\operatorname{tr}(W'W)^p} = \frac{1}{K} = \ell_k^*.$$

Thus, in Proposition 3

$$\Sigma^* = \frac{K}{4} \frac{(W'W)^p}{\operatorname{tr}(W'W)^p},$$

as required.

For the optimal IV, consider the thin singular value decomposition that drops the zero eigenvalues:

$$W = U\Lambda V'$$

where Λ is diagonal with strictly positive elements and U and V have orthonormal columns. Thus, $W'W = V\Lambda^2 V'$, $(W'W)^p = V\Lambda^{2p}V'$, and

$$\Sigma^* = \frac{K}{4} \frac{V\Lambda^{2p}V'}{\operatorname{tr}((W'W)^p)}.$$

Moreover,

$$S = W\Sigma^* W' = \frac{K}{4} \frac{U\Lambda^{2p+2}U'}{\operatorname{tr}((W'W)^p)}.$$

Thus,

$$\left(S^\dagger\right)^{1/(p+1)} = \left(\frac{K}{4\operatorname{tr}((W'W)^p)}\right)^{-1/(p+1)} U\Lambda^{-2}U' = aU\Lambda^{-2}U',$$

and the Theorem 1 optimal IV is

$$\begin{aligned}
\left(S^\dagger\right)^{1/(p+1)} \tilde{x} &= \left(S^\dagger\right)^{1/(p+1)} W\tilde{g} \\
&= aU\Lambda^{-1}V'\tilde{g} \\
&= a\left(W'\right)^\dagger \tilde{g}.
\end{aligned}$$

When W is block-diagonal, the relaxed problem can be solved block-by-block, implying that Σ^* is the same. Moreover, since the implied S is block-diagonal, Theorem 1 also applies block-by-block.

A.6 Proof of Theorem 2 and Associated Claims

In this section, we first prove Theorem 2. Then, we provide brief proofs of claims made in the main text about sufficient conditions for Assumption 1. Finally, we prove that $\hat{v} \xrightarrow{p} v^*$. Define $\mathcal{A}_K$ as the event under which Assumption 1 holds. Next, define the good event $\mathcal{G}_K = \mathcal{A}_K \cap \{v_K > 0\}$. By Assumption 1, $\mathbb{P}(\mathcal{G}_K) \to 1$. For the convergence-in-probability and convergence-in-distribution statements in the proof, which are not affected by modifying vanishing-in-probability sequences, we therefore work without loss of generality as if $\mathcal{G}_K$ holds. We start by decomposing the estimator error as:

$$\sqrt{K}(\widehat{\beta} - \beta) = \frac{K^{-1/2}b'\tilde{g}}{K^{-1}\tilde{g}'B\left(\frac{1}{2}\mathbf{1} + \tilde{g}\right)}, \tag{A13}$$

where for this section we use the shorthand notation $\tilde{g} = g^{GR} - 0.5 \cdot \mathbf{1}$ for the recentered Gaussian rounded shocks, S instead of S_{GR} and $B = W'(S^{\dagger})^{1/(p+1)}W$. The proof of the theorem then follows by combining the decomposition in Equation (A13) with Lemma A2 and Lemma A3 by Slutsky's theorem.

We handle the denominator and numerator separately. First, for the numerator, we use an existing result on dependency graph central limit theorems (CLTs) from Chen and Shao (2004, Theorem 2.4).[36] The denominator law of large numbers (LLN) relies on a more direct argument that also relies on sparsity in Σ^*.

Lemma A2. *Under Assumption 1,*

$$K^{-1/2} \sum_{k=1}^{K} b_k \tilde{g}_k \Rightarrow \mathcal{N}(0, v^*).$$

Proof. We first prove a CLT conditional on a sequence of W and ε. We write $Y_k = K^{-1/2} b_k \tilde{g}_k$, so that

$$\operatorname{Var}\left[\sum_k Y_k \mid W, \varepsilon\right] = \frac{1}{K} b' \Sigma^{GR} b = v_K.$$

Conditional on W and ε, Y is an element-wise transformation of ξ, which is a Gaussian random vector. This means that independence properties pass from ξ to Y and so if $\Sigma^*_{kl} = 0$ then Y_k and Y_l are conditionally independent. Let $\mathcal{G}$ be the dependency graph for $Y|W, \varepsilon$ that has K vertices, with an edge between vertex k and $l \neq k$ if Y_k and Y_l are dependent, conditional on (W, ε); equivalently, there is an edge if and only if $\Sigma^*_{kl} \neq 0$ and $k \neq l$.

Define the s-neighborhood of k as $N^s[k] = \{l : \operatorname{dist}_{\mathcal{G}}(k, l) \leq s\}$, where the distance measure is the shortest path between two vertices, and we use the convention that $d_{\mathcal{G}}(k, k) = 0$. $N^1[k]$, $N^2[k]$ and $N^3[k]$ are three sets that play the role of the sets A_i, B_i, C_i from condition LD3 in Chen and Shao (2004). Specifically, conditionally on W, ε:

- Y_k is independent of any Y_t with $t \notin N^1[k]$;
- Y_l for any $l \in N^1[k]$ is independent of any Y_t with $t \notin N^2[k]$ because $N^2[k]$ contains all neighbors of those in $N^1[k]$;
- Y_l for any $l \in N^2[k]$ is independent of any Y_t with $t \notin N^3[k]$.

The next step in applying the theorem is computing an upper bound, for any k, on:

$$\begin{aligned}
&\max\{|\{l : N^3[k] \cap N^2[l] \neq \emptyset\}|, |N^3[k]|\} \\
&\leq \max\{|N^5[k]|, |N^3[k]|\} \\
&\leq (\bar{d} + 1)^5.
\end{aligned}$$

[36]This result would allow us to grow the degree of Σ^* slowly, where the rate of growth has to be slower for higher p, but to keep things simple we assume that the degree does not grow with K.

Next, we define the conditional normal approximation error for the centered and normalized random variable $\sum_{k=1}^{K} Y_k/\sqrt{v_K}$:

$$\delta_K = \sup_z \left| \mathbb{P}\left\{ \frac{K^{-1/2}\sum_{k=1}^{K} b_k \tilde{g}_k}{\sqrt{v_K}} \leq z \mid W, \varepsilon \right\} - \Phi(z) \right|.$$

We can now apply Theorem 2.4 of Chen and Shao (2004), which gives us the first line of the below display.

$$\begin{aligned}
\delta_K &\leq 75(\bar{d}+1)^{10} \sum_{k=1}^{K} \mathbb{E}\left[|Y_k/\sqrt{v_K}|^3 \mid W, \varepsilon \right] \\
&\leq 75(\bar{d}+1)^{10} \frac{K^{-3/2}}{8 v_K^{3/2}} \sum_{k=1}^{K} |b_k|^3 = o_p(1).
\end{aligned}$$

The second line uses the fact that $|\tilde{g}_k| = 1/2$, recalling that $Y_k = K^{-1/2} b_k \tilde{g}_k$, and Assumption 1(e). We now have a conditional CLT; by the bounded convergence theorem, since $0 \leq \delta_K \leq 1$, we have that $\mathbb{E}[\delta_K] \to 0$ and we also have an unconditional CLT. □

We finish by providing an LLN for the denominator of the estimator.

Lemma A3. *Under Assumption 1,*

$$K^{-1} \tilde{g}' B \left(\tfrac{1}{2}\mathbf{1} + \tilde{g} \right) \xrightarrow{p} h^*.$$

Proof. The expectation of the first term is zero, and

$$\mathbb{E}\left[\tilde{g}' B \tilde{g} \mid W \right] = \operatorname{tr}(B\Sigma^{GR}) = \operatorname{tr}\left\{ (S^{\dagger})^{1/(p+1)} S \right\} = \operatorname{tr}(S^{p/(p+1)}),$$

so $K^{-1}\mathbb{E}\left[\tilde{g}' B \left(\frac{1}{2}\mathbf{1} + \tilde{g} \right) \mid W \right] = h_K$. By Assumption 1(c), $h_K \xrightarrow{p} h^*$. It remains to prove that

$$K^{-1} \tilde{g}' B \left(\tfrac{1}{2}\mathbf{1} + \tilde{g} \right) - h_K \xrightarrow{p} 0,$$

which will be accomplished by showing that the variance of the term goes to zero as $K \to \infty$. The arcsin transformation does not change sparsity, so the row-sum sparsity of Σ^* passes to Σ^{GR}. Since $|\Sigma^{GR}_{kl}| \leq 1/4$ for all k, l, this means that every row has absolute sum of at most $(\bar{d}+1)/4$; hence

$$\left\| \Sigma^{GR} \right\|_\infty \leq (\bar{d}+1)/4, \qquad \|S\|_\infty \leq \|W\|_\infty^2 \left\| \Sigma^{GR} \right\|_\infty \leq (\bar{d}+1)\bar{\lambda}/4,$$

where the last inequality uses the second part of Assumption 1(a). Furthermore, expanding definitions, we have that:

$$B\Sigma^{GR}B = W'(S^{\dagger})^{1/(p+1)}S(S^{\dagger})^{1/(p+1)}W = W'S^{(p-1)/(p+1)}W$$

So,

$$\begin{aligned}
\mathrm{Var}\left[K^{-1}\tilde{g}'B(\tfrac{1}{2}\mathbf{1}) \mid W\right] &= K^{-2}\big(W(\tfrac{1}{2}\mathbf{1})\big)'S^{(p-1)/(p+1)}\big(W(\tfrac{1}{2}\mathbf{1})\big) \\
&\leq K^{-2}\left\|S^{(p-1)/(p+1)}\right\|_{\infty}\|W(\tfrac{1}{2}\mathbf{1})\|_2^2 \\
&\leq K^{-2}\|S\|_{\infty}^{(p-1)/(p+1)}\|W\|_{\infty}^2 K/4 \\
&\leq C_1/K,
\end{aligned}$$

for a finite constant C_1. For the quadratic term, by Assumption 1(a,b), $\|B\|_{\infty} \leq \|W\|_{\infty}^2 \left\|(S^{\dagger})^{1/(p+1)}\right\|_{\infty} = \lambda_{\max}(W'W)\{\lambda_{\min}^{+}(S)\}^{-1/(p+1)} \leq \bar{\lambda}\cdot\underline{\lambda}^{-1/(p+1)}$. Then, using Lemma A4, applied conditionally on W, and the fact that $\mathrm{rank}(B) \leq K$, we have:

$$\mathrm{Var}\left[K^{-1}\tilde{g}'B\tilde{g} \mid W\right] \leq CK^{-2}(\bar{d}+1)^2\|B\|_2^2 \leq CK^{-2}(\bar{d}+1)^2\,\mathrm{rank}(B)\,\|B\|_{\infty}^2 \leq C_2/K$$

for a finite constant C_2. Using $\mathrm{Var}[X+Y \mid W] \leq 2\mathrm{Var}[X \mid W] + 2\mathrm{Var}[Y \mid W]$, we get

$$\mathrm{Var}\left[K^{-1}\tilde{g}'B\left(\tfrac{1}{2}\mathbf{1}+\tilde{g}\right)\middle| W\right] \leq 2(C_1+C_2)/K.$$

For every $\eta > 0$, Chebyshev's inequality gives

$$\mathbb{P}\left(\left|K^{-1}\tilde{g}'B\left(\tfrac{1}{2}\mathbf{1}+\tilde{g}\right) - h_K\right| > \eta \,\middle|\, W\right) \leq \frac{C}{\eta^2 K}.$$

Taking expectations yields

$$\mathbb{P}\left(\left|K^{-1}\tilde{g}'B\left(\tfrac{1}{2}\mathbf{1}+\tilde{g}\right) - h_K\right| > \eta\right) \leq \frac{C}{\eta^2 K} \to 0.$$

Therefore $K^{-1}\tilde{g}'B\left(\tfrac{1}{2}\mathbf{1}+\tilde{g}\right) - h_K \xrightarrow{p} 0$. □

Lemma A4. *Let $\tilde{g}_k \in \{-1/2, 1/2\}$ be centered random variables with a dependency graph whose maximum degree is at most $\bar{d}$. Then, for every symmetric and fixed $K \times K$ matrix B,*

$$\mathrm{Var}\left[g'\tilde{B}\tilde{g}\right] \leq C(\bar{d}+1)^2\|B\|_2^2,$$

where C is a finite constant.

Proof. Because $\tilde{g}_k^2 = 1/4$, we can write

$$\tilde{g}'B\tilde{g} = \frac{1}{4}\mathrm{tr}(B) + Q, \qquad Q = 2\sum_{k<l} B_{kl}\tilde{g}_k\tilde{g}_l.$$

Hence $\text{Var}\left[\tilde{g}' B \tilde{g}\right] = \text{Var}\left[Q\right]$, and

$$\text{Var}\left[Q\right] = 4 \sum_{k<l} \sum_{a<b} B_{kl} B_{ab} \text{Cov}\left[\tilde{g}_k \tilde{g}_l, \tilde{g}_a \tilde{g}_b\right].$$

We next count for fixed $k < l$, how many pairs $a < b$ can have nonzero covariance with $\tilde{g}_k \tilde{g}_l$. As before, let $\mathcal{G}$ be the dependency graph of $\tilde{g}$ and let $N^s[k] = \{l : \text{dist}_{\mathcal{G}}(k, l) \leq s\}$. To have non-zero covariance, at least one of a or b must be in $N^1[k] \cup N^1[l]$, i.e. $\{a, b\} \cap \left(N^1[k] \cup N^1[l]\right) \neq \emptyset$. After choosing one such endpoint, say $a \in N^1[k] \cup N^1[l]$, the other endpoint must satisfy $b \in N^1[k] \cup N^1[l] \cup N^1[a]$; otherwise the covariance is zero. So, the number of non-zero covariance partners of the kl pair is at most:

$$2\left|N^1[k] \cup N^1[l]\right| \max_{a \in N^1[k] \cup N^1[l]} \left|N^1[k] \cup N^1[l] \cup N^1[a]\right| \leq 12(\bar{d}+1)^2.$$

Since $|\tilde{g}_k \tilde{g}_l| \leq 1/4$, all covariance terms are bounded by a universal constant. Using the count above, the symmetry of the covariance relation, and $|xy| \leq (x^2 + y^2)/2$,

$$\begin{aligned}
\text{Var}\left[Q\right] &\leq C \sum_{k<l,\ a<b:\, \text{Cov}[\tilde{g}_k \tilde{g}_l, \tilde{g}_a \tilde{g}_b] \neq 0} |B_{kl} B_{ab}| \\
&\leq C \sum_{k<l,\ a<b:\, \text{Cov}[\tilde{g}_k \tilde{g}_l, \tilde{g}_a \tilde{g}_b] \neq 0} (B_{kl}^2 + B_{ab}^2) \\
&\leq C(\bar{d}+1)^2 \sum_{k<l} B_{kl}^2 \leq C(\bar{d}+1)^2 \left\|B\right\|_2^2,
\end{aligned}$$

where the constant C can change from line to line. Thus $\text{Var}\left[\tilde{g}' B \tilde{g}\right] \leq C(\bar{d}+1)^2 \left\|B\right\|_2^2$. □

We next provide brief proofs of some claims made in the discussion of Assumption 1(a) in the main text.

Proposition A1. *Under the assumptions of Proposition 3, when p is an integer, the Proposition 3 maximizer Σ^* of the relaxed problem (7), and therefore its Gaussian rounding implementation Σ^{GR}, satisfy $d(\Sigma^*) = d(\Sigma^{GR}) \leq (d(W'W))^p$.*

Proof. From Proposition 3, $\Sigma^* = \frac{1}{4} D_{\ell^*}^{-1/2} \frac{M_{\ell^*}^p}{\text{tr}\left(M_{\ell^*}^p\right)} D_{\ell^*}^{-1/2}$, where $M_\ell = D_\ell^{-1/2} W' W D_\ell^{-1/2}$, D_ℓ is a non-singular diagonal matrix, and $\text{tr}\left(M_{\ell^*}^p\right)$ is a non-zero scalar. Thus we can write

$$\Sigma^* \propto D_{\ell^*}^{-1}\left(W'W\right) D_{\ell^*}^{-1}\left(W'W\right) \cdots D_{\ell^*}^{-1},$$

with p copies of $W'W$, or equivalently

$$\Sigma_{kl}^* \propto \sum_{t_1=1}^{K} \cdots \sum_{t_{p-1}=1}^{K} \frac{(W'W)_{kt_1} \cdot (W'W)_{t_1 t_2} \cdot \ldots \cdot (W'W)_{t_{p-1} l}}{D_{kk} \cdot D_{t_1 t_1} \cdot \ldots \cdot D_{ll}}.$$

Therefore, the non-zero entries of Σ^* require at least one path through p non-zero entries of $W'W$.

There are $(d(W'W))^p$ such paths. Since the non-zero elements are the same for Σ^* and Σ^{GR}, $d(\Sigma^*) = d(\Sigma^{GR})$, so the inequality holds for $d(\Sigma^{GR})$ as well. $\square$

For the next claim in the main text, if W has row and column sparsity bounded by constants c and r respectively, then $d(W'W) \leq cr$ and $(d(W'W))^p$ are also bounded as long as p is finite, implying the first part of Assumption 1(a). If the entries of W are also bounded, then row and column sparsity means that row and column sums of the absolute values of the entries of W are bounded. The largest eigenvalue of $W'W$, $\lambda_{\max}(W'W)$, which is the spectral norm of W, squared, is bounded by the product of the maximum absolute column sum and maximum absolute row sum of W. This yields the second part of Assumption 1(a).

We next show $K \asymp N$ in non-degenerate cases. Suppose we have both row and column sparsity, and a non-vanishing fraction of outcome units are connected to at least some shocks as K grows, meaning there exist constants $\rho > 0$ and $\underline{w} > 0$ such that, with probability $1 - o(1)$,

$$\frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left\{ \sum_{k=1}^{K} W_{ik}^2 \geq \underline{w} \right\} \geq \rho.$$

On this event,

$$\begin{aligned} \rho \underline{w} N &\leq \sum_{i=1}^{N} \sum_{k=1}^{K} W_{ik}^2 \\ &= \operatorname{tr}(W'W) \\ &\leq K \lambda_{\max}(W'W) \\ &\leq \bar{\lambda} K. \end{aligned}$$

This implies that we cannot have $N/K \to \infty$.

Conversely, we have already shown that for a finite constant C, $\lambda_{\max}(S) \leq C$. So,

$$\begin{aligned} h_K &= \frac{1}{K} \operatorname{tr}\left(S^{p/(p+1)}\right) \\ &\leq \frac{\operatorname{rank}(S)}{K} \lambda_{\max}(S_{GR})^{p/(p+1)} \\ &\leq C^{p/(p+1)} \frac{N}{K}, \end{aligned}$$

So $h_K \xrightarrow{p} 0$ if $N/K \to 0$, contradicting Assumption 1(c). Therefore, $K \asymp N$.

Finally, we prove asymptotic validity of our confidence intervals constructed using $\hat{v}$, as defined in the main text.

Proposition A2. *Under the assumptions of Theorem 2, $\hat{v} \xrightarrow{p} v^*$.*

Proof. We can write $\hat{v} = v_K - 2(\widehat{\beta} - \beta)K^{-1}x'M\varepsilon + (\widehat{\beta} - \beta)^2 K^{-1} x'Mx$, where $M = S_{GR}^{(p-1)/(p+1)}$. By Theorem 2, $\hat{\beta} - \beta = o_p(1)$ and, by Assumption 1, $v_K - v^* = o_p(1)$. To finish the proof, we need that $a_K = K^{-1}x'M\varepsilon$ and $\omega_K^2 = K^{-1}x'Mx$ are both $O_p(1)$. First, write $\omega_K^2 \leq K^{-1} \|M\|_\infty \lambda_{\max}(W'W) \|g\|^2 = O(1)$. This is because $\|M\|_\infty = \lambda_{\max}(S)^{(p-1)/(p+1)}$ and, by the argument in Lemma A3, $\lambda_{\max}(S)$ is bounded, as well as $\|g\|^2 \leq K$. Second, by Cauchy-Schwartz, since $M \succeq 0$, $|a_K| \leq K^{-1}\sqrt{(x'Mx)(\varepsilon' M\varepsilon)} = \omega_K \sqrt{v_K} = O_p(1)$. $\square$

B Additional Results

B.1 Interpretation with Heterogeneous Effects

Suppose the true outcome model is

$$y_i = \beta_i x_i + \varepsilon_i,$$

where β_i now represents a heterogeneous treatment effect for unit i. For any design δ, consider any partially-whitened instrument from Theorem 1, $z^* = z^*_\delta$ with $z^*_i = \left(S_\delta^{\dagger/(p+1)}\tilde{x}\right)_i$ for any p. Following Borusyak and Hull (2024, 2026), we characterize the IV estimand, i.e., ratio of $\mathbb{E}_\delta\left[\sum_i z_i y_i\right]$ and $\mathbb{E}_\delta\left[\sum_i z_i x_i\right]$. Treating (β_i, ε_i) as fixed without loss of generality:

$$\begin{aligned}\frac{\mathbb{E}_\delta\left[\sum_i z^*_i y_i\right]}{\mathbb{E}_\delta\left[\sum_i z^*_i x_i\right]} &= \frac{\mathbb{E}_\delta\left[\sum_i z^*_i \beta_i x_i\right]}{\mathbb{E}_\delta\left[\sum_i z^*_i x_i\right]} \\ &= \frac{\sum_i \mathrm{Cov}_\delta\left[z^*_i, x_i\right]\beta_i}{\sum_i \mathrm{Cov}_\delta\left[z^*_i, x_i\right]},\end{aligned}$$

which is a convex average of heterogeneous effects when $\mathrm{Cov}_\delta\left[z^*_i, x_i\right] \geq 0$ for all i. To see that this holds, write:

$$\mathrm{Cov}_\delta\left[z^*, x\right] = \mathbb{E}_\delta\left[z^*\tilde{x}'\right] = S_\delta^{\dagger/(p+1)} S_\delta = S_\delta^{p/(p+1)}.$$

This matrix is positive semidefinite so every diagonal element is nonnegative. These convex weights are known given the known design δ and exposure mapping of x.[37]

In the model with both direct and spillover effects, this result generally breaks down due to contamination bias (Goldsmith-Pinkham et al., 2024) except in special cases. Specifically, rewrite (5) with heterogeneous effects model as:

$$y_i = \beta_i x_i + \tau_i g_i + \varepsilon_i,$$

and note that since the instrument z^* from Theorem 1$'$ is orthogonal to g we can write the estimand for the spillover effect as:

$$\frac{\mathbb{E}_\delta\left[\sum_i z^*_i y_i\right]}{\mathbb{E}_\delta\left[\sum_i z^*_i x_i\right]} = \frac{\sum_i \mathrm{Cov}_\delta\left[z^*_i, x_i\right]\beta_i}{\sum_i \mathrm{Cov}_\delta\left[z^*_i, x_i\right]} + \frac{\sum_i \mathbb{E}_\delta\left[z^*_i \tilde{g}_i\right]\tau_i}{\sum_i \mathrm{Cov}_\delta\left[z^*_i, x_i\right]}$$

[37]We note, however, that the convex weighting result does not generally extend to the nonlinear outcome model $y_i = y_i(x_i)$, except when the optimal instrument is unwhitened. The partially whitened IV estimate identifies a weighted average of causal effects $\partial y_i(x)/\partial x$ at different margins x with weights proportional to $\sum_j (S_\delta^{\dagger/(p+1)})_{ij}\mathrm{Cov}_\delta\left[x_j, \mathbf{1}[x_i \geq x] \mid y(\cdot)\right]$ (see Appendix S.1 of Borusyak and Hull (2026)) which are not guaranteed to be non-negative.

The first term is a weighted average of β_i, as before, but the weights need not be convex.[38] Moreover, the second term is not zero since we do not generally have $\mathbb{E}_\delta\left[z_i^* \tilde{g}_i\right] = 0$ for each i: just $\mathbb{E}_\delta\left[\sum_i z_i^* \tilde{g}_i\right] = 0$. An exception is when the direct effects are constant, $\tau_i = \tau$, or otherwise uncorrelated with $\mathbb{E}_\delta\left[z_i^* \tilde{g}_i\right]$.

B.2 Non-Recentered Estimators

This appendix extends the main results to allow for non-recentered instruments; specifically, to characterize the designs and instruments which minimize a finite-sample approximate worst-case mean-squared error (MSE) allowing for some bias. We show that the MSE-optimal design and estimator have the same form as the baseline design and estimator with recentering, but that the optimal design can have undesirable properties. We further show that the MSE advantage of non-recentered instruments disappears when appropriate covariates are included.

Consider an extension of the Borusyak and Hull (2026) approximation, now to approximate mean-squared error (MSE) of an IV estimator using a possibly non-recentered instrument $z_i(g)$:

$$\mathcal{M}_{\delta,\mathcal{E}}\left[z\right] = \frac{\mathbb{E}_{\delta,\mathcal{E}}\left[\left(\sum_{i=1}^N z_i(g)\varepsilon_i\right)^2\right]}{\mathbb{E}_\delta\left[\sum_{i=1}^N z_i(g) x_i\right]^2}.$$

Following the same steps as in the Appendix A.1–A.2 proofs of Theorem 1 and Proposition 1, one can show the MSE-optimal design and estimator

$$\delta^*, z^* \in \arg\min_{\delta\in\mathcal{D}, z} \max_{\mathcal{E}\in\mathcal{F}_p(\sigma)} \mathcal{M}_{\delta,\mathcal{E}}\left[z\right]$$

are, for $M_\delta = \mathbb{E}_\delta\left[xx'\right]$:

$$\delta^* \in \arg\max_{\delta\in\mathcal{D}} \operatorname{tr}\left(M_\delta^{\frac{p}{p+1}}\right),$$
$$z^* \propto \left(M_{\delta^*}^\dagger\right)^{\frac{1}{p+1}} x.$$

The MSE-optimal instrument is generally not recentered (i.e., $\mathbb{E}_{\delta^*}\left[z_i^*(g)\right] \neq 0$ in general), implying some small bias can help for this problem.

However, the optimal design can have undesirable properties. To see this simply, consider the no-spillovers special case where $W = I$. Previously, the optimal design was independent randomization with $\mathbb{E}_{\delta^*}\left[g_i\right] = 0.5$ for all i. Now we have the following:

Proposition A3. *When $W = I$ and $p < \infty$, an optimal design $\delta^* \in \arg\max_{\delta\in\mathcal{D}} \operatorname{tr}\left(M_\delta^{\frac{p}{p+1}}\right)$ is*

[38] Appendix A.1 of Goldsmith-Pinkham et al. (2024) shows convexity is guaranteed in an OLS regression with binary x_i, since then the relevant weight matrix is proportional to an M-matrix. The analogous sufficient condition here is that the instrument is whitened and that the generally non-binary x_i still satisfies $\left|\mathrm{Cov}_\delta\left[x_i, g_i\right]\right| \leq \min\{\mathrm{Var}_\delta\left[x_i\right], \mathrm{Var}_\delta\left[g_i\right]\}$.

given by complete randomization with K_1 treated units, where

$$K_1 = \begin{cases} \lfloor m^* \rfloor, & w.p.\, 1-\theta^* \\ \lfloor m^* \rfloor + 1, & w.p.\, \theta^* \end{cases}$$

for

$$\begin{aligned} m^* &\in \arg\max_{m\in[0,N]} \left[\left(\frac{v(m)}{N}\right)^{\frac{p}{p+1}} + (N-1)\left(\frac{Nm - v(m)}{N(N-1)}\right)^{\frac{p}{p+1}} \right], \\ v(m) &= \lfloor m \rfloor^2 + (2\lfloor m \rfloor + 1)(m - \lfloor m \rfloor), \\ \theta^* &= m^* - \lfloor m^* \rfloor. \end{aligned}$$

In the case where m^* is an integer, this result shows complete randomization with exactly $K_1 = m^*$ units treated is MSE-optimal. The expression for m^* shows that when $p = 1$ the optimal share of treated units solves

$$\max_{t\in[0,1]} t + \sqrt{N-1} \cdot \sqrt{t(1-t)}$$

giving $t^* = (1 + 1/\sqrt{N})/2 > 1/2$. As $p \to \infty$ with N fixed, this treated share grows to one. Intuitively, without recentering, the constant mean direction counts towards the "signal" of the treatment leading to unusually high treatment rates.

A complementary result shows that the MSE advantage of non-recentered estimators disappears (i.e., there is no MSE cost to recentering) whenever $W\mathbf{1}$ is spanned by controls (included as in Appendix B.8, below). In the previous no-spillovers special case, this would hold when estimation includes an intercept:

Proposition A4. *Let R be a matrix of predetermined controls, let $M_R = I - R(R'R)^\dagger R'$ and assume $W\mathbf{1} \in Col(R)$. Then there is no loss in only considering recentered instruments when minimizing worst-case MSE; i.e. when solving*

$$\delta^*, z^* \in \arg\min_{\delta\in\mathcal{D}, z\in\mathcal{Z}} \max_{\mathcal{E}\in\mathcal{F}_p(\sigma)} \frac{\mathbb{E}_{\delta,\mathcal{E}}\left[(z(g)'M_R\varepsilon)^2\right]}{\mathbb{E}_\delta\left[z(g)'M_R x\right]^2},$$

it is without loss to limit to z with $\mathbb{E}_\delta[z_i(g)] = 0$ for all i.

Intuitively, with an intercept, the non-recentered instrument can no longer leverage the constant mean direction which can yield unusually high treatment rates (while allowing for some bias). Once an intercept or, more generally, the expected treatment $W\mathbf{1}$ is absorbed by the controls there is no reason to not recenter.

B.3 Equicorrelated ε_i

Consider $\varepsilon \in \mathbb{R}^N$ where, for some $\nu > 0$ and $\rho \in [-(N-1)^{-1}, 1]$,

$$\Omega_\rho = \mathbb{E}\left[\varepsilon\varepsilon'\right] = \nu^2\left[(1-\rho)I + \rho\mathbf{1}\mathbf{1}'\right].$$

Here $\nu^2 = \mathrm{Var}\left[\varepsilon_i\right]$ and $\rho = \mathrm{Corr}\left[\varepsilon_i, \varepsilon_j\right]$ for $j \neq i$. Since $\Omega_\rho = \nu^2\left[(1-\rho)(I - \mathbf{1}\mathbf{1}'/N) + (1+(N-1)\rho)\mathbf{1}\mathbf{1}'/N\right]$ by Lemma A5:

$$\|\Omega_\rho\|_p = \left[\frac{1}{N}(1+(N-1)\rho)^p + \frac{N-1}{N}(1-\rho)^p\right]^{1/p} \cdot \nu^2 N^{1/p}.$$

Thus, $\Omega_\rho \in \mathcal{F}_p(\sigma)$ for some $\sigma \geq \nu$ if and only if

$$\left[\frac{1}{N}(1+(N-1)\rho)^p + \frac{N-1}{N}(1-\rho)^p\right]^{1/p} \leq \frac{\sigma^2}{\nu^2}. \tag{A14}$$

This condition is illustrated in Figure A3 which shows the maximum ν^2 corresponding to each ρ, depending on $N \in \{10, 50\}$ and $p \in \{1, 2, \infty\}$. It is evident that smaller p makes the condition weaker: For $p = 1$, it holds for any ρ and N. But for larger p, the condition penalizes correlations, especially in large samples: ν^2 has to be much smaller to stay within $\mathcal{F}_p(\sigma)$.

B.4 Rank-1 Designs

Proposition A5. *When $p = \infty$, there is always an optimal design δ^* supported on a single pair of complementary assignments g^* and $\mathbf{1} - g^*$ so $\mathrm{Var}_\delta[g]$ has rank 1. Moreover, when $W \geq 0$ elementwise, one such design corresponds to setting $g = \mathbf{1}$ with probability 0.5 and $g = \mathbf{0}$ otherwise.*

B.5 Symmetry Normalization

Proposition A6. *Suppose $\mathcal{G}$ is a finite group of relabelings of observations and shocks represented by pairs of permutation matrices (P, Q) such that $PWQ' = W$ for every $(P, Q) \in \mathcal{G}$. Then there exists a solution δ^* to Proposition 1 such that $\mathbb{E}_{\delta^*}[g_k] = 0.5$, for all k and δ^* is $\mathcal{G}$-invariant: i.e., if $g \sim \delta^*$ then Qg has the same distribution as g for every $(P, Q) \in \mathcal{G}$. Equivalently, δ^* may be chosen to assign equal probability to assignments in the same $\mathcal{G}$-orbit, i.e. for any $g, g' \in \{0,1\}^K$ such that $g' = Qg$ for some $(P, Q) \in \mathcal{G}$.*

In particular, this result shows that if W is invariant under permutations of a subset of shocks (possibly together with corresponding permutations of observations), then we may without loss restrict attention to designs that are exchangeable within that subset. Similarly, if W has multiple isomorphic blocks, we may take the induced block-level design to be exchangeable across those blocks. A version of this result also goes through for the multiple-exposure extension, with W replaced by $W - cU$, as long as $W - cU$ is invariant under $\mathcal{G}$ for every c.

B.6 Special Case: Leave-Out Averages

No Direct Effects. In this special case the intervention units coincide with outcome units, which are partitioned into J clusters C_j of sizes N_j with $x_i = \bar{g}_{-i} = \frac{1}{N_{j(i)}-1}\sum_{k\in C_{j(i)},k\neq i} g_k$. We know from Proposition 2 that it is optimal to randomize across clusters with $\mathbb{E}_\delta[g_i] = 0.5$. Focusing on cluster j, the within-cluster exposure matrix can be written as

$$W_{(j)} = \frac{\mathbf{1}_{N_j}\mathbf{1}_{N_j}' - I_{N_j}}{N_j - 1} = -\frac{1}{N_j - 1}(I_{N_j} - P_{N_j}) + P_{N_j},$$

with

$$W_{(j)}'W_{(j)} = (N_j - 1)^{-2}(I_{N_j} - P_{N_j}) + P_{N_j}.$$

We solve the relaxed problem for cluster j and propose a feasible implementation of this solution. Since $W_{(j)}$ is symmetric, we apply Corollary 2 to get

$$\Sigma_{(j)}^* = \frac{1}{4}\frac{\left(W_{(j)}'W_{(j)}\right)^p}{\operatorname{tr}\left(\left(W_{(j)}'W_{(j)}\right)^p\right)/N_j}.$$

By Lemma A5,

$$\left(W_{(j)}'W_{(j)}\right)^p = r_j\left(I_{N_j} - P_{N_j}\right) + P_{N_j}$$

and

$$\operatorname{tr}\left(\left(W_{(j)}'W_{(j)}\right)^p\right) = r_j(N_j - 1) + 1,$$

such that the off-diagonal elements of the within-cluster correlation matrix $4\Sigma_{(j)}^*$ are all equal to

$$\psi_j = \frac{(1 - r_j)/N_j}{r_j(N_j - 1)/N_j + 1/N_j} = \frac{1 - r_j}{1 + (N_j - 1)r_j},$$

Since $r_j \in [0, 1]$ and therefore $\psi_j \in [0, 1]$, this relaxed problem solution can be implemented by mixing clustered assignment and the Bernoulli design with probabilities ψ_j and $1 - \psi_j$.

For the optimal instrument, for $i \in C_j$ write

$$\tilde{x}_i = x_i - 0.5 = \frac{N_j\bar{g}_j - g_i}{N_j - 1} - 0.5 = (\bar{g}_j - 0.5) + \frac{\bar{g}_j - g_i}{N_j - 1}, \tag{A15}$$

where the first term is cluster-level and the second term averages to zero within the cluster. Moreover,

$$S_{(j)} = W_{(j)}\Sigma_{(j)}^*W_{(j)} \propto \frac{W_{(j)}\left(W_{(j)}'W_{(j)}\right)^p W_{(j)}}{\operatorname{tr}\left(\left(W_{(j)}'W_{(j)}\right)^p\right)/N_j} = \frac{(N_j - 1)^{-2(p+1)}\left(I_{N_j} - P_{N_j}\right) + P_{N_j}}{r_j(N_j - 1)/N_j + 1/N_j}$$

and, applying Lemma A5 again,

$$\begin{aligned}S_{(j)}^{-1/(p+1)} &= \left(r_j(N_j-1)/N_j + 1/N_j\right)^{1/(p+1)} \cdot \left((N_j-1)^2\left(I_{N_j} - P_{N_j}\right) + P_{N_j}\right) \\ &= \left(1 + (N_j-1)\psi_j\right)^{-1/(p+1)} \cdot \left((N_j-1)^2\left(I_{N_j} - P_{N_j}\right) + P_{N_j}\right).\end{aligned}$$

Moreover, for $\tilde{x}_{(j)} = W_{(j)}\tilde{g}_{(j)}$ where $\tilde{g}_{(j)}$ is the appropriate subvector of $\tilde{g}$,

$$\begin{aligned}z_i^* &= \left(S_{(j)}^{-1/(p+1)}\tilde{x}_{(j)}\right)_i \\ &= \left(S_{(j)}^{-1/(p+1)}W\tilde{g}_{(j)}\right)_i \\ &= \left(1 + (N_j-1)\psi_j\right)^{-1/(p+1)} \cdot \left(-(N_j-1)\left(I_{N_j} - P_{N_j}\right)\tilde{g}_{(j)} + P_{N_j}\tilde{g}_{(j)}\right)_i \\ &= \left(1 + (N_j-1)\psi_j\right)^{-1/(p+1)}\left((\bar{g}_j - 0.5) + (N_j-1)\left(\bar{g}_j - g_i\right)\right) \cdot\end{aligned}$$

With Direct Effects. We start again by characterizing the solution of the relaxed problem. For a fixed c,

$$\begin{aligned}W_{(j),c} = W_{(j)} - cI_{N_j} &= \frac{1}{N_j-1}\left(\mathbf{1}_{N_j}\mathbf{1}_{N_j}' - \left(1 + c(N_j-1)\right)I_{N_j}\right) \\ &= -\left(\frac{1}{N_j-1} + c\right)\left(I_{N_j} - P_{N_j}\right) + (1-c)P_{N_j}.\end{aligned}$$

Thus,

$$W_{(j),c}'W_{(j),c} = \left(\frac{1}{N_j-1} + c\right)^2\left(I_{N_j} - P_{N_j}\right) + (1-c)^2 P_{N_j}$$

and, by Lemma A5,

$$\left(W_{(j),c}'W_{(j),c}\right)^p = \left|\frac{1}{N_j-1} + c\right|^{2p}\left(I_{N_j} - P_{N_j}\right) + |1-c|^{2p} P_{N_j}$$

with

$$\operatorname{tr}\left(\left(W_{(j),c}'W_{(j),c}\right)^p\right) = \left|\frac{1}{N_j-1} + c\right|^{2p}(N_j-1) + |1-c|^{2p}.$$

By Corollary 2, the correlation matrix is given by

$$4\Sigma_{(j)}^* = \frac{\left(W_{(j),c}'W_{(j),c}\right)^p}{\operatorname{tr}\left(\left(W_{(j),c}'W_{(j),c}\right)^p\right)/N_j} = \frac{\left|\frac{1}{N_j-1} + c\right|^{2p}\left(I_{N_j} - P_{N_j}\right) + |1-c|^{2p} P_{N_j}}{\frac{N_j-1}{N_j}\left|\frac{1}{N_j-1} + c\right|^{2p} + \frac{1}{N_j}|1-c|^{2p}}$$

with off-diagonal elements

$$\psi_j^{\perp}(c) = \frac{|1-c|^{2p} - \left|\frac{1}{N_j-1} + c\right|^{2p}}{|1-c|^{2p} + (N_j-1)\left|\frac{1}{N_j-1} + c\right|^{2p}} = \frac{1 - r_j^{\perp}(c)}{1 + (N_j-1)r_j^{\perp}(c)}$$

for $r_j^{\perp}(c) = \left(\frac{\left|\frac{1}{N_j-1}+c\right|}{|1-c|}\right)^{2p}$. Thus we have

$$S_{(j),c} = W_{(j),c}\Sigma^*_{(j)}W'_{(j),c} \propto \frac{\left|\frac{1}{N_j-1}+c\right|^{2(p+1)}(I_{N_j}-P_{N_j}) + |1-c|^{2(p+1)}P_{N_j}}{\frac{N_j-1}{N_j}\left|\frac{1}{N_j-1}+c\right|^{2p}+\frac{1}{N_j}|1-c|^{2p}}.$$

Then c^* corresponding to the optimal design minimizes

$$\begin{aligned}\sum_j \operatorname{tr}\left(S_{(j),c}^{p/(p+1)}\right) &= \sum_j N_j^{p/(p+1)} \frac{\operatorname{tr}\left(\left|\frac{1}{N_j-1}+c\right|^{2p}(I_{N_j}-P_{N_j}) + |1-c|^{2p}P_{N_j}\right)}{\left((N_j-1)\left|\frac{1}{N_j-1}+c\right|^{2p}+|1-c|^{2p}\right)^{p/(p+1)}} \\ &= \sum_j N_j^{p/(p+1)}\left((N_j-1)\left|\frac{1}{N_j-1}+c\right|^{2p}+|1-c|^{2p}\right)^{1/(p+1)}. \qquad \text{(A16)}\end{aligned}$$

In general, we cannot guarantee that $\psi_j^{\perp}(c^*) \geq 0$. However, for even N_j, any such $\Sigma^*_{(j)}$ is implementable by mixing cluster-level shocks with complete randomization within the cluster.[39]

To characterize the optimal spillover instrument, we have

$$\begin{aligned}S_{(j),c}^{-1/(p+1)} \propto &\left(\frac{N_j-1}{N_j}\left|\frac{1}{N_j-1}+c\right|^{2p}+\frac{1}{N_j}|1-c|^{2p}\right)^{1/(p+1)} \\ &\cdot\left(\left(\frac{1}{N_j-1}+c\right)^{\dagger 2}(I_{N_j}-P_{N_j}) + (1-c)^{\dagger 2}P_{N_j}\right),\end{aligned}$$

where $x^{\dagger} = x^{-1}$ if $x \neq 0$ and 0 otherwise for scalar x, and thus

$$\begin{aligned} z_i^* &\propto \left(S_{(j),c}^{-1/(p+1)}W_{(j),c}\tilde{g}_{(j)}\right)_i \\ &\propto \left(\frac{N_j-1}{N_j}\left|\frac{1}{N_j-1}+c\right|^{2p}+\frac{1}{N_j}|1-c|^{2p}\right)^{1/(p+1)} \cdot \left(-\left(\frac{1}{N_j-1}+c\right)^{\dagger}(I_{N_j}-P_{N_j})\tilde{g}_{(j)} + (1-c)^{\dagger}P_{N_j}\tilde{g}_{(j)}\right)_i \\ &= \left(\frac{N_j-1}{N_j}\left|\frac{1}{N_j-1}+c\right|^{2p}+\frac{1}{N_j}|1-c|^{2p}\right)^{1/(p+1)} \cdot \left((1-c)^{\dagger}(\bar{g}_j-0.5) + \left(\frac{1}{N_j-1}+c\right)^{\dagger}(\bar{g}_j-g_i)\right),\end{aligned}$$

evaluated at $c = c^*$.

We now consider two special cases. First, when all clusters are of the same size, $N_j = n$, the

[39] This is not exactly true for odd N_j where the most negative correlation implementable by binary shocks is $-1/N_j$ (compared to the $-1/(N_j-1)$ bound guaranteed by positive semi-definiteness). While we are not sure whether $\psi_j^{\perp}(c^*) < -1/N_j$ is possible, if it is, we conjecture that the best design is one that achieves correlations $-1/N_j$ by randomly choosing whether to treat $(N_j-1)/2$ or $(N_j+1)/2$ units in the cluster and then randomly choosing which ones.

optimal design simplifies:

$$c^* = \arg\min_c (n-1) \left| \tfrac{1}{n-1} + c \right|^{2p} + |1-c|^{2p},$$

with the first order condition that implies

$$\frac{c^* + \frac{1}{n-1}}{1-c^*} = (n-1)^{-1/(2p-1)}.$$

Plugging this in gives $r^\perp(c^*) = (n-1)^{-2p/(2p-1)} \in [0,1]$ with a common optimal mixing probability of $\psi^\perp(c^*) = (1 - r^\perp(c^*))/(1 + (n-1)r^\perp(c^*)) \in [0,1]$. Thus, the design is implementable by mixing cluster-level assignment with Bernoulli randomization. Noting that $c^* \notin \{-1/(n-1), 1\}$, the optimal spillover instrument simplifies to

$$z_i^* \propto \left(\bar{g}_{j(i)} - \frac{1}{2}\right) + (n-1)^{1/(2p-1)}(\bar{g}_{j(i)} - g_i).$$

Second, when $p = 1$ (but cluster sizes can vary), the first-order condition in (A16) yields $c^* = 0$. Thus, the design and optimal IV are the same as in the setting without direct effects.

B.7 Special Case: Spatial Spillovers

Fix $p = 1$ and assume for simplicity that $2L+1$ divides N and $2L+1 \leq N/2$. That is, the circle can be split into an integer number of blocks of length $2L+1$, at least two of them.

We first solve the relaxed problem. The structure of spillovers implies $\|w_{\cdot k}\|_2 = 1/\sqrt{2L+1}$ and $w_{\cdot k}' w_{\cdot l} = \frac{1}{(2L+1)^2} \max\{2L+1-d(k,l), 0\}$ for $k, l \in \{1, \ldots, N\}$, where $d(k,l)$ is the circular distance between units. Thus, by Corollary 1 the covariances in the relaxed problem's solution,

$$\Sigma_{kl}^* = \frac{1}{4} \max\left\{1 - \frac{d(k,l)}{2L+1}, 0\right\},$$

decay linearly reaching zero at distance $2L+1$. This covariance matrix can be implemented with binary shocks, specifically by the procedure described in the main text. Indeed, the correlation between g_k and g_l equals the probability that they are in the same block, which equals $\max\left\{1 - \frac{d(k,l)}{2L+1}, 0\right\}$, as required.

For this symmetric design, Corollary 2 implies $z^* \propto W^\dagger \tilde{g}$. We now verify the characterization of z^* in the main text. First, since an L-neighborhood of any unit i includes the center of the block to which i belongs and no other block center, averaging the values of z^* within a neighborhood yields $\frac{1}{2L+1}\left((2L+1)\,\tilde{g}_i - 2L\bar{\tilde{g}} + 2L\bar{\tilde{g}}\right) = \tilde{g}_i$. Thus, $Wz^* = \tilde{g}$. For this z^* to be $W^\dagger \tilde{g}$, it remains to show that $z^* \perp \ker(W)$, i.e. $z^{*\prime} v = 0$ for any v such that $Wv = 0$. The structure of W implies that its kernel consists of vectors v that are $(2L+1)$-periodic with zero sums in each block: $v_{i+(2L+1)} = v_i$ and $\sum_{i=1}^{2L+1} v_i = 0$. Moreover, the structure of z^* implies that its sums over any $(2L+1)$-spaced orbit (i.e., set of units with the same position relative to the block center) is the same, $\frac{N}{2L+1}\bar{\tilde{g}}$. Thus,

$\sum_{i=1}^{N} z_i^* v_i = \sum_{i=1}^{2L+1} \frac{N}{2L+1} \bar{\bar{g}} v_i = 0$, which concludes the proof. We note that, despite $p = 1$, this optimal IV is not fully whitened because S is singular in this setting.

B.8 Predetermined Covariates

This appendix extends the optimal design and instrument results to specifications with predetermined controls. We allow the model error to contain an arbitrary component in the span of the controls and show that the controlled problem is exactly the baseline problem after residualizing exposures and instruments: i.e., all formulas apply with W replaced by $M_R W$.

Let $R = (r_i')$ be an $N \times L$ matrix of predetermined covariates and let $M_R = I - R(R'R)^\dagger R'$ be the associated residual-maker matrix. We expand the class of allowed error distributions to include linear combinations of r_i with unrestricted coefficients:

$$\mathcal{F}_p^R(\sigma) = \left\{ \mathcal{E} : \varepsilon_i = \gamma' r_i + \tilde{\varepsilon}_i,\, \gamma \in \mathbb{R}^L,\, \left\| \mathbb{E}_{\mathcal{E}} \left[\tilde{\varepsilon} \tilde{\varepsilon}' \right] \right\|_p \leq N^{1/p} \sigma^2 \right\}.$$

Let

$$\tilde{x}_{\delta,R} = M_R W \tilde{g}_\delta,$$
$$S_{\delta,R} = \mathrm{Var}_\delta \left[\tilde{x}_{\delta,R} \right] = M_R W \Sigma_\delta W' M_R.$$

We then have the following results:

Proposition A7. *Fix δ and suppose $S_{\delta,R} \neq 0$. Then the minimax optimal recentered instrument is:*

$$z^*_{\delta,R}(g) = \left(S^\dagger_{\delta,R} \right)^{\frac{1}{p+1}} \tilde{x}_{\delta,R} \in \arg\min_{z \in \mathcal{Z}_\delta} \max_{\mathcal{E} \in \mathcal{F}_p^R(\sigma)} \mathcal{V}_{\delta,\mathcal{E}} \left[z \right],$$

with the worst-case approximate variance of

$$\max_{\mathcal{E} \in \mathcal{F}_p^R(\sigma)} \mathcal{V}_{\delta,\mathcal{E}} \left[z^*_{\delta,R} \right] = N^{1/p} \sigma^2 \operatorname{tr}(S_{\delta,R}^{p/(p+1)})^{-(p+1)/p}.$$

Proposition A8. *Suppose there is $\delta \in \mathcal{D}$ such that $S_{\delta,R} \neq 0$. Then any*

$$\delta_R^* \in \arg\max_{\delta \in \mathcal{D}} \operatorname{tr}(S_{\delta,R}^{p/(p+1)})$$

together with the instrument from Proposition A7 solves the minimax problem.

Proposition A9. *There exists a maximizer δ_R^* from Proposition A8 such that $\mathbb{E}_{\delta_R^*} \left[g_k \right] = 0.5$, for all k. Moreover, if $M_R W$ is block-diagonal after a joint partition of rows and columns, then there exists such a δ_R^* with independent subvectors of g across blocks.*

The proofs here follow by showing the R-controlled problem is equivalent to the original problem with W replaced by the columnwise residualized matrix $M_R W$. Notice that the optimal instrument

$z^*_{\delta,R}(g)$ is always in-sample orthogonal to R; variation in directions absorbed by R is not rewarded by the objective and so is not used. A sufficient condition for block-diagonality of $M_R W$ is that M_R and W are jointly block-diagonal after the same joint partition of rows and columns, such as when all covariates are included with interactions with the dummies of blocks in W.

B.9 Nonlinear Spillover Formulas

It is straightforward to extend Theorem 1 and Proposition 1 to settings where $x_i = X_i(g)$ for a nonlinear (but still known) set of formulas $X(\cdot) = (X_1(\cdot), \dots, X_N(\cdot))'$ which can depend on W, as studied in Borusyak and Hull (2023). Indeed, the Appendix A.1-A.2 proofs extend verbatim after defining the general $\tilde{x}_\delta = x - \mathbb{E}_\delta[x]$ and $S_\delta = \mathrm{Var}_\delta[x]$. The central challenge with this extension is computation: unlike in the linear case, the optimal design objective is not reducible to a search over covariance matrices of the g shocks and is thus generally much more difficult. Still, a researcher can compute and maximize this objective over a relatively small set of candidate designs $\mathcal{D}_r \subset \mathcal{D}$:

$$\max_{\delta \in \mathcal{D}_r} \mathrm{tr}\left(\mathrm{Var}_\delta[X(g)]^{p/(p+1)}\right).$$

The second and third claims of Proposition 2 extend similarly.

However, the first claim of Proposition 2 — that there is an optimal design with $\mathbb{E}_\delta[g_k] = 0.5$ — does not generalize. We illustrate this and build intuition for nonlinear solutions in a special case in which $N = K$ observations are clustered into groups C_j and $x_i = \mathbf{1}\left[\sum_{k \in C_{j(i)}, k \neq i} g_k \geq 1\right]$ indicates that i has at least one shocked neighbor in his group. Without loss, consider a design which optimizes group-by-group and is exchangeable within groups. Focusing on one cluster and dropping the j subscript for brevity, let $T = \sum_{i \in C} g_i$ and

$$\begin{aligned} a &= Pr_\delta(T = 0) \\ b &= Pr_\delta(T = 1) \\ 1 - a - b &= Pr_\delta(T \geq 2), \end{aligned}$$

with the uniquely shocked unit chosen at random when $T = 1$. We then have the following result:

Proposition A10. *Fix $n = |C| \geq 3$ and $p \in [1, \infty)$ in the above nonlinear spillover formula case. Then among exchangeable designs,*

$$\max_\delta \mathrm{tr}(S_\delta^{p/(p+1)})$$

has a unique solution where $Pr_\delta(T = 0) = a^*$ and $Pr_\delta(T = 1) = 1 - a^*$ for a^* solving

$$1 - 2a^* = (a^*)^{1/(p+1)}(n-1)^{-(p-1)/(p+1)}.$$

Hence the optimal design never shocks more than one unit in the group, and $\mathbb{E}_\delta[g_k] \neq 1/2$. Moreover:

(a) If $n = 3$ or $p = 1$ then $a^* = 0.25$;

(b) As $p \to \infty$, $a^* \to \frac{n-2}{2(n-1)}$;

(c) As $n \to \infty$ *for* $p > 1$, $a^* \to 1/2$.

Intuitively, the optimal design mixes between shocking nobody in a cluster or shocking exactly one unit since any additional shocks are redundant in terms of the nonlinear spillover treatment. As isotropy becomes less important ($p \to \infty$), the mixing probability approaches the value which maximizes the total variance of the x_i.[40] As group sizes get large ($n \to \infty$), the mixing probability approaches 1/2 because the identity of the shocked unit becomes second-order and we approach the group-specific shock special case.

B.10 Budget Constraints

For the optimal design, consider first the analytically simpler case where the researcher faces a constraint that the probability of treating each unit be exactly q. Proposition 1 is then modified to

$$\delta^* \in \arg \max_{\delta:\, \mathbb{E}_\delta[g]=q\mathbf{1}} \operatorname{tr}\left\{ \left(W \operatorname{Var}_\delta [g]\, W' \right)^{p/(p+1)} \right\}.$$

In the case of a block-diagonal W, this problem still splits into block-by-block parts, so this claim of Proposition 2 extends, too.

We propose to formulate the relaxed problem as an optimization over the set of all correlation matrices multiplied by $q(1-q)$ instead of 1/4. Clearly, the objective $\operatorname{tr}(W\Sigma W')^{p/(p+1)}$ is just a rescaling relative to $\Sigma \in \mathcal{Q}_K$, and the optimal covariance matrix Σ^* is the same, rescaled by $4q(1-q)$. The Theorem 1 characterization of the relaxed problem's solution is therefore unchanged, too. The Gaussian rounding approximation, in turn, is modified to drawing $\xi \sim \mathcal{N}(0, \Sigma^*/\left(q(1-q)\right))$ and setting $g_k = \mathbf{1}\left[\xi_k \geq \Phi^{-1}(1-q)\right]$ where $\Phi(\cdot)$ is the standard normal CDF. Since $\operatorname{Var}[\xi_k] = 1$, we have $\mathbb{E}_\delta[g_k] = q$ for all k as required.[41]

We now consider three alternative constraints the researcher may face. First, suppose the budget constraint takes the inequality form $\mathbb{E}_\delta[g_k] \leq q$. In general, we do not know whether the analog of the first claim of Proposition 2 holds, i.e. that all of the inequality constraints have to bind, with $\mathbb{E}_\delta[g_k] = q$. This is true, however, in the relaxed problem, as for any $\Sigma \succeq 0$ with $\Sigma_{kk} \leq q(1-q)$ and some inequalities strict, we can consider $\tilde{\Sigma} = \Sigma + \operatorname{diag}(q(1-q) - \Sigma_{kk}) \succeq \Sigma$ which yields a weakly lower worst-case variance.

Second, suppose the budget constraint is soft, in the sense that the researcher is willing to trade off the worst-case estimation variance with the expected cost of treating more units. For instance,

[40] One can extend the result to $p = \infty$; as before, the optimal exchangeable design is not unique and here it does not generally set $Pr_\delta(T > 1) = 0$.

[41] Note that the optimization problem could impose an additional restriction that $\Sigma^*_{kl} \geq -\min\left\{q^2, (1-q)^2\right\}$ for $k \neq l$, which has to hold for any binary vector with marginals q. We do not recommend this, however, as the Gaussian rounding procedure already ensures this inequality starting from any $\Sigma^* \succeq 0$. Imposing extra constraints on Σ^* would unnecessarily limit the set of Gaussian rounding designs.

one can choose q by solving

$$\min_{q\in[0,1/2]}\left\{\left(\max_{\delta:\,\mathbb{E}_\delta[g]=q\mathbf{1}}\operatorname{tr}\left(\left(W\mathrm{Var}_\delta\left[g\right]W'\right)^{p/(p+1)}\right)\right)^{-(p+1)/p}+\lambda qK\right\},$$

where the first term is the worst-case variance of the estimator and the second term is the expected cost of the experiment, with weight λ.[42] This is a straightforward scalar optimization, and it is especially simple if the inner optimization is replaced by its relaxed version, as the optimal shock correlation matrix does not depend on q.

The final scenario is when the researcher's constraint is on the realized number of treated units: $\frac{1}{K}\sum_k g_k\leq q$. This formulation is difficult to implement within our approach, as it does not naturally translate to a restriction on Σ and it is not consistent with Gaussian rounding. In cases like this, we recommend using our original constraint on $\mathbb{E}_\delta\left[g_k\right]$ but either with a tighter q (such that it is very unlikely that the realization violates the constraint on the realized number of treated units, as formalized in Sun (2026)) or with small *ad hoc* modifications (e.g., switching a randomly chosen set of $g_k=1$ to $g_k=0$ until the constraint holds).

B.11 Error Bound on the Gaussian Rounding Implementation

Proposition A11. *We have:*

$$\operatorname{tr}\left(S_{GR}^{p/(p+1)}\right)\geq\left(\frac{2}{\pi}\right)^{p/(p+1)}\max_{\Sigma\in\mathcal{C}_K}\operatorname{tr}\left(\left(W\Sigma W'\right)^{p/(p+1)}\right).$$

Thus, the worst-case approximate variance of the Theorem 1 optimal recentered IV estimator under the Gaussian rounding design based on the relaxed problem's solution from Proposition 3 is at most $\pi/2$ times that of the original Proposition 1 problem. The same result holds for the multiple-exposure extension in Sections 2.4 and 3.1.

C Proofs of Additional Results

C.1 A Useful Lemma

For $n\in\mathbb{N}$, denote $P_n=\mathbf{1}_n\mathbf{1}_n'/n$.

Lemma A5. *Consider $n\in\mathbb{N}$ and $\alpha,\beta,\gamma\in\mathbb{R}$ such that $\alpha,\beta\geq 0$ and if $\gamma<0$ then $\alpha,\beta>0$. Then the matrix $A=\alpha(I_n-P_n)+\beta P_n$ is positive semi-definite and we have:*

$$A^\gamma=\alpha^\gamma(I_n-P_n)+\beta^\gamma P_n.$$

Moreover,

$$\operatorname{tr}\left(A^\gamma\right)=(n-1)\alpha^\gamma+\beta^\gamma$$

[42] Note the $q\in[0,1/2]$ domain is without loss, as it is never optimal to pick $q>1/2$.

and, when $\gamma \geq 1$,

$$\|A\|_\gamma = ((n-1)\alpha^\gamma + \beta^\gamma)^{1/\gamma}.$$

Proof. The Lemma statement provides a spectral decomposition for matrix A, since P_n is a rank 1 orthogonal projector onto the span of $\mathbf{1}_n$ and $I_n - P_n$ is a rank $n-1$ orthogonal projector onto the space orthogonal to $\mathbf{1}_n$. This means that A has $n-1$ eigenvalues that are α and one eigenvalue that is β, and the rest of the claims come from spectral properties of the trace and the Schatten-γ norm. □

C.2 Proposition A3

With no spillovers, the design problem is simply:

$$\max_\delta \operatorname{tr}\left(\mathbb{E}_\delta \left[gg'\right]^{\frac{p}{p+1}}\right).$$

By symmetry and concavity of $A \mapsto \operatorname{tr}(A^{p/(p+1)})$, there is an exchangeable optimizer. Any such design can be represented by first drawing $K_1 = \mathbf{1}'g$ from some distribution and then choosing K_1 treated units at random. For such a design, $\mathbb{E}_\delta\left[g_k^2\right] = \mathbb{E}_\delta\left[g_k\right] = \mathbb{E}_\delta\left[\frac{K_1}{N}\right]$ and, for $k \neq l$,

$$\mathbb{E}_\delta\left[g_k g_\ell\right] = \mathbb{E}_\delta\left[\mathbb{E}_\delta\left[g_k g_\ell \mid K_1\right]\right] = \mathbb{E}_\delta\left[\frac{K_1(K_1-1)}{N(N-1)}\right].$$

Simple algebra yields $\mathbb{E}_\delta\left[gg'\right] = \lambda_{1,\delta} P_N + \lambda_{2,\delta}\left(I_N - P_N\right)$ for

$$\lambda_{1,\delta} = \frac{\mathbb{E}_\delta\left[K_1^2\right]}{N}, \qquad \lambda_{2,\delta} = \frac{\mathbb{E}_\delta\left[K_1(N-K_1)\right]}{N(N-1)}.$$

By Lemma A5, the objective is:

$$\max_\delta \lambda_{1,\delta}^{p/(p+1)} + (N-1)\lambda_{2,\delta}^{p/(p+1)}.$$

Now fix $m = \mathbb{E}_\delta\left[K_1\right] \in [0, N]$. We have $\lambda_{1,\delta} \geq \lambda_{2,\delta}$ since $\lambda_{1,\delta} - \lambda_{2,\delta} = \mathbb{E}_\delta\left[K_1(K_1-1)\right]/(N-1) \geq 0$, with equality only if $K_1 \in \{0, 1\}$ a.s., i.e. $\mathbb{E}_\delta\left[K_1^2\right] = m$. Moreover, the derivative of the objective with respect to $\mathbb{E}_\delta\left[K_1^2\right]$ is

$$\frac{p}{p+1} \cdot \left(\frac{1}{N}\lambda_{1,\delta}^{-1/(p+1)} - \frac{1}{N}\lambda_{2,\delta}^{-1/(p+1)}\right) \leq 0,$$

with strict inequality except at a single point $\mathbb{E}_\delta\left[K_1^2\right] = m$. Thus, the objective is strictly decreasing in $\mathbb{E}_\delta\left[K_1^2\right]$. Hence, for fixed m, we minimize $\mathbb{E}_\delta\left[K_1^2\right]$. That is achieved by a distribution for K_1 that has support only on the two adjacent integers around m. Specifically, for $k = \lfloor m \rfloor$ and $\theta = m - k \in [0, 1)$, it should place probability $1-\theta$ on k and probability θ on $k+1$, such that

$$\mathbb{E}_\delta\left[K_1^2\right] = v(m) = (1-\theta)k^2 + \theta(k+1)^2 = k^2 + (2k+1)(m-k).$$

The optimal m is given by maximizing the objective:

$$m^* \in \arg\max_{m\in[0,N]} \left[\left(\frac{v(m)}{N}\right)^{p/(p+1)} + (N-1)\left(\frac{Nm - v(m)}{N(N-1)}\right)^{p/(p+1)}\right],$$

which yields the desired solution.

C.3 Proposition A4

It is straightforward to show (by the same steps as above) that for fixed δ the optimal IV is

$$z^*_\delta(g) \propto (A^\dagger_\delta)^{1/(p+1)} M_R W g$$

where

$$A_\delta = M_R W \mathbb{E}_\delta\left[gg'\right] W' M_R$$

and that optimal designs using this instrument are given by

$$\delta^* \in \arg\max_\delta \operatorname{tr}(A_\delta^{p/(p+1)}).$$

We now show it is without loss in this problem to restrict to designs with equal marginals. For any δ, let δ^c be the complement design and let $\delta^{sym} = \frac{1}{2}\delta + \frac{1}{2}\delta^c$. Since $x^c = W(\mathbf{1} - g) = W\mathbf{1} - x$ and $W\mathbf{1} \in \mathrm{Col}(R)$, we have

$$M_R x^c = M_R(W\mathbf{1} - x) = -M_R x.$$

Therefore

$$\begin{aligned} A_{\delta^c} &= M_R \mathbb{E}_{\delta^c}\left[xx'\right] M_R \\ &= \mathbb{E}_\delta\left[(M_R x^c)(M_R x^c)'\right] \\ &= \mathbb{E}_\delta\left[(M_R x)(M_R x)'\right] \\ &= A_\delta, \end{aligned}$$

and also $A_{\delta^{sym}} = \frac{1}{2}A_\delta + \frac{1}{2}A_{\delta^c} = A_\delta$. Hence restricting to symmetrized designs is without loss, and they all have equal marginals.

Finally, note that under any equal-marginal design,

$$M_R W(g - \mathbb{E}_\delta[g]) = M_R W(g - \frac{1}{2}\mathbf{1}) = M_R W g.$$

Therefore the optimal instrument can be written

$$z_\delta^*(g) \propto (A_\delta^\dagger)^{1/(p+1)} M_R W(g - \mathbb{E}_\delta[g])$$

which is recentered. Therefore, it is without loss to restrict attention to recentered instruments.

C.4 Proposition A5

When $p = \infty$, the problem in Proposition 1 becomes

$$\max_\delta \operatorname{tr}(S_\delta) = \max_\delta \operatorname{tr}(W \operatorname{Var}_\delta[g] W').$$

Moreover, by Proposition 2 we may restrict attention to designs with $\mathbb{E}_\delta[g] = \frac{1}{2}\mathbf{1}$. Define $h = 2g - 1 \in \{-1, 1\}^K$ where $\mathbb{E}_\delta[h] = 0$ and $\operatorname{Var}_\delta[g] = \frac{1}{4}\mathbb{E}_\delta[hh']$. Hence,

$$\operatorname{tr}(S_\delta) = \operatorname{tr}(W \operatorname{Var}_\delta[g] W') = \operatorname{tr}(W'W \operatorname{Var}_\delta[g]) = \frac{1}{4}\mathbb{E}_\delta\left[h'W'Wh\right].$$

Choose h^* to maximize $h'W'Wh$ over $\{-1,1\}^K$ and let δ^* put equal probability on $g^* = \frac{1}{2}(\mathbf{1} + h^*)$ and $1 - g^* = \frac{1}{2}(\mathbf{1} - h^*)$. Here $\mathbb{E}_{\delta^*}[g] = \frac{1}{2}\mathbf{1}$ and

$$\operatorname{tr}(S_{\delta^*}) = \frac{1}{4}h^{*\prime}W'Wh^* \geq \frac{1}{4}\mathbb{E}_\delta\left[h'W'Wh\right] = \operatorname{tr}(S_\delta)$$

for every design δ with $\mathbb{E}_\delta[g] = \frac{1}{2}\mathbf{1}$. Therefore δ^* is optimal. Finally, note that

$$\operatorname{Var}_{\delta^*}[g] = \left(g^* - \frac{1}{2}\mathbf{1}\right)\left(g^* - \frac{1}{2}\mathbf{1}\right)' = \frac{1}{4}h^*h^{*\prime}$$

has rank 1. Moreover, when $W \geq 0$, $W'W \geq 0$ so $h'W'Wh$ is maximized by $h = \pm\mathbf{1}$; i.e., setting all $g_k = 1$ with probability 0.5 and otherwise setting all $g_k = 0$.

C.5 Proposition A6

By Proposition 2, we may restrict attention to designs with $\mathbb{E}_\delta[g_k] = 0.5$. For any $(P, Q) \in \mathcal{G}$, let $\delta^{(P,Q)}$ be the distribution of Qg for $g \sim \delta$. Since Q is a relabeling of shocks, each $\delta^{(P,Q)}$ also satisfies $\mathbb{E}_{\delta^{(P,Q)}}[g] = \frac{1}{2}\mathbf{1}$. We have $\Sigma_{\delta^{(P,Q)}} = Q\Sigma_\delta Q'$. Moreover, since Q is a relabeling matrix, $Q'Q = I$ such that $PWQ' = W$ implies $PW = WQ$. Thus,

$$S_{\delta^{(P,Q)}} = W\Sigma_{\delta^{(P,Q)}}W' = WQ\Sigma_\delta Q'W' = PW\Sigma_\delta W'P' = PS_\delta P'.$$

Since P is an orthogonal matrix, $\operatorname{tr}\left(S_{\delta^{(P,Q)}}^{p/(p+1)}\right) = \operatorname{tr}\left(S_{\delta}^{p/(p+1)}\right)$. Now average δ over the transformations in $\mathcal{G}$, defining

$$\bar{\delta} = |\mathcal{G}|^{-1} \sum_{(P,Q)\in\mathcal{G}} \delta^{(P,Q)}.$$

Because all designs in this average have the same mean, $\Sigma_{\bar{\delta}} = |\mathcal{G}|^{-1} \sum_{(P,Q)\in\mathcal{G}} Q\Sigma_\delta Q'$ and $S_{\bar{\delta}} = |\mathcal{G}|^{-1} \sum_{(P,Q)\in\mathcal{G}} PS_\delta P'$. And since $A \mapsto \operatorname{tr}(A^{p/(p+1)})$ is weakly concave on the positive semidefinite cone for $p \in [1, \infty]$,

$$\operatorname{tr}\left(S_{\bar{\delta}}^{p/(p+1)}\right) \geq |\mathcal{G}|^{-1} \sum_{(P,Q)\in\mathcal{G}} \operatorname{tr}\left(\left(PS_\delta P'\right)^{p/(p+1)}\right) = \operatorname{tr}\left(S_{\delta}^{p/(p+1)}\right).$$

Thus averaging over the symmetry group weakly improves the Proposition 1 objective while producing a $\mathcal{G}$-invariant design. Applying this argument to an optimal δ^* proves the main result.

It remains only to justify the final equivalence. For the $\mathcal{G}$-invariant optimal design δ^* and for $a \in \{0,1\}^K$ write $\delta^*(a) = \Pr_{\delta^*}[g = a]$. For every $(P,Q) \in \mathcal{G}$, if $g \sim \delta^*$ then $Qg \overset{d}{=} g$ (where $\overset{d}{=}$ stands for equal in distribution). Take any two assignments $a, b \in \{0,1\}^K$ in the same $\mathcal{G}$-orbit. Then there exists $(P,Q) \in \mathcal{G}$ such that $b = Qa$. By $\mathcal{G}$-invariance,

$$\delta^*(b) = \Pr_{\delta^*}[g = b] = \Pr_{\delta^*}[Qg = b] = \Pr_{\delta^*}\left[g = Q^{-1}b\right] = \Pr_{\delta^*}[g = a] = \delta^*(a).$$

Hence δ^* is constant on each $\mathcal{G}$-orbit, i.e. it assigns equal probability to assignments in the same orbit.

Conversely, if a design δ is constant on each $\mathcal{G}$-orbit, then for any $(P,Q) \in \mathcal{G}$ and any $A \subseteq \{0,1\}^K$,

$$\Pr_\delta[Qg \in A] = \sum_{a\in A} \Pr_\delta[Qg = a] = \sum_{a\in A} \Pr_\delta\left[g = Q^{-1}a\right] = \sum_{a\in A} \delta(Q^{-1}a) = \sum_{a\in A} \delta(a) = \Pr_\delta[g \in A],$$

because $Q^{-1}a$ and a are in the same $\mathcal{G}$-orbit. Thus $Qg \overset{d}{=} g$, proving equivalence.

C.6 Propositions A7, A8, and A9

The following facts about the projection matrix are useful: M_R is symmetric and idempotent, $M_R R = 0$, $\|M_R\|_\infty \leq 1$, and for a vector v, $M_R v = v$ if and only if $R'v = 0$.

For Proposition A7, we first show that any $z \in \mathcal{Z}_\delta$ with $\max_{\mathcal{E}\in\mathcal{F}_p^R(\sigma)} \mathcal{V}_{\delta,\mathcal{E}}[z] < \infty$ has to satisfy $R'z = 0$ a.s. Indeed, for $\gamma \in \mathbb{R}^L$ let $\mathcal{E}_\gamma \in \mathcal{F}_p^R(\sigma)$ put probability one on $\varepsilon = R\gamma$, such that

$$\mathcal{V}_{\delta,\mathcal{E}_\gamma}[z] \propto \operatorname{Var}_\delta\left[\gamma' R' z\right] = \gamma' \operatorname{Var}_\delta\left[R'z\right]\gamma.$$

If $\operatorname{Var}_\delta[R'z] \neq 0$, we can pick γ to make $\mathcal{V}_{\delta,\mathcal{E}_\gamma}[z]$ arbitrarily large, such that $\sup_{\mathcal{E}\in\mathcal{F}_p^R(\sigma)} \mathcal{V}_{\delta,\mathcal{E}}[z] \geq \sup_\gamma \mathcal{V}_{\delta,\mathcal{E}_\gamma}[z] = \infty$. Thus, $\operatorname{Var}_\delta[R'z] = 0$ and, since $\mathbb{E}_\delta[R'z] = 0$, we have $R'z = 0$ a.s.

Thus, there is a solution to $\inf_{z \in \mathcal{Z}_\delta} \max_{\mathcal{E} \in \mathcal{F}_p^R(\sigma)} \mathcal{V}_{\delta,\mathcal{E}}[z]$ that belongs to $\mathcal{Z}_\delta^R = \{z \in \mathcal{Z}_\delta \colon R'z = 0 \text{ a.s.}\}$, so we will now look for that solution. Let $W_R = M_R W$, $x_R = W_R g$, $\tilde{x}_R = W_R \tilde{g}$, and $q = p/(p-1)$, and note that for $z \in \mathcal{Z}_\delta^R$ and $\varepsilon = R\gamma + \tilde{\varepsilon}$ we have $z'\varepsilon = z'\tilde{\varepsilon}$. Moreover, using $z = M_R z$ and $\mathbb{E}_\delta[z] = 0$ for the denominator,

$$\mathcal{V}_{\delta,\mathcal{E}}[z] = \frac{\mathrm{Var}_{\delta,\mathcal{E}}[z'\varepsilon]}{\mathbb{E}_\delta[z'Wg]^2} = \frac{\mathrm{Var}_{\delta,\mathcal{E}}[z'\tilde{\varepsilon}]}{\mathbb{E}_\delta[z'W_R\tilde{g}]^2} = \frac{\mathrm{Var}_{\delta,\mathcal{E}}[z'\tilde{\varepsilon}]}{\mathbb{E}_\delta[z'\tilde{x}_R]^2}.$$

Thus, recalling that $Q_z = \mathbb{E}_\delta[zz']$, analogously to (A1),

$$\max_{\mathcal{E} \in \mathcal{F}_p^R(\sigma)} \mathcal{V}_{\delta,\mathcal{E}}[z] = \max_{\tilde{\mathcal{E}} \in \mathcal{F}_p(\sigma)} \frac{\mathrm{Var}_{\delta,\tilde{\mathcal{E}}}[z'\tilde{\varepsilon}]}{\mathbb{E}_\delta[z'\tilde{x}_R]^2} = N^{1/p}\sigma^2 \frac{\|Q_z\|_q}{\mathbb{E}_\delta[z'\tilde{x}_R]^2}. \tag{A17}$$

Our goal is to find $z \in \mathcal{Z}_\delta^R$ that minimizes the right-hand side of (A17). Minimizing that function over the larger set $\mathcal{Z}_\delta$ is the Theorem 1 optimization problem with the exposure matrix W_R instead of W, which has the desired solution $z_{\delta,R}^*$ and the worst-case variance given in the proposition. It is therefore the solution over $\mathcal{Z}_\delta^R$ as long as $z_{\delta,R}^* \in \mathcal{Z}_\delta^R$. To see this, note that $S_{\delta,R} = M_R W \Sigma_\delta W' M_R = M_R S_{\delta,R} M_R$ and therefore $\mathrm{Col}(S_{\delta,R}) \subseteq \mathrm{Col}(M_R)$. Moreover, $\left(S_{\delta,R}^\dagger\right)^{1/(p+1)}$ has the same column space as $S_{\delta,R}$ (including for $p = \infty$ under our convention that $\left(S_{\delta,R}^\dagger\right)^0 = S_{\delta,R}S_{\delta,R}^\dagger$). It follows that, for every g, $z_{\delta,R}^* = \left(S_{\delta,R}^\dagger\right)^{1/(p+1)} \tilde{x}_R \in \mathrm{Col}(M_R)$ and therefore $M_R z_{\delta,R}^* = z_{\delta,R}^*$ a.s. or, equivalently, $R' z_{\delta,R}^* = 0$ a.s., completing the proof.

Similarly, applying Proposition 1 to the same residualized problem proves Proposition A8, while Proposition A9 follows from Proposition 2.

C.7 Proposition A10

Let $u = Pr_\delta(x_i = 0) = a + \frac{b}{n}$, which does not depend on i by exchangeability. Then:

$$\begin{aligned} \mathbb{E}_\delta[x_i] &= 1 - u \\ \mathrm{Var}_\delta[x_i] &= u(1-u) \\ \mathrm{Cov}_\delta[x_i, x_j] &= a - u^2, \qquad i \neq j \end{aligned}$$

Thus,

$$S_\delta = (u-a)(I_n - P_n) + (nu(1-u) - (n-1)(u-a))\,P_n$$

and, by Lemma A5,

$$\mathrm{tr}\left(S_\delta^{p/(p+1)}\right) = (n-1)(u-a)^{p/(p+1)} + (nu(1-u) - (n-1)(u-a))^{p/(p+1)}.$$

Letting $\lambda = u - a$, we can rewrite the objective function as

$$\Phi_u(\lambda) = (n-1)\lambda^{p/(p+1)} + (nu(1-u) - (n-1)\lambda)^{p/(p+1)}.$$

Since $p/(p+1) \in (0,1)$, $\Phi_u(\cdot)$ is concave and from the first order condition is maximized at $\lambda = u(1-u)$. But feasibility requires

$$\begin{aligned} 1 - a - b = 1 - u - (n-1)\lambda \geq 0 \\ \implies \lambda \leq \frac{1-u}{n-1}. \end{aligned}$$

Hence for fixed u the maximizer is

$$\lambda^*(u) = \min\left\{u(1-u), \frac{1-u}{n-1}\right\}.$$

Consider first the case $u \leq 1/(n-1)$. Then the $u(1-u)$ solution is feasible and

$$\operatorname{tr}(S_\delta^{p/(p+1)}) = n(u(1-u))^{p/(p+1)},$$

which is increasing in u on $[0, 1/(n-1)]$. Hence the optimum over this region is attained on the boundary $u = 1/(n-1)$. Thus, without loss we can restrict attention to the other case, $u \geq 1/(n-1)$ and $\lambda = (1-u)/(n-1)$. In this case, $b^* = 1 - a^*$ such that $Pr_{\delta^*}(T \geq 2) = 0$. Moreover, $u = a + \frac{1-a}{n}$ and the objective simplifies to

$$\operatorname{tr}(S_\delta^{p/(p+1)}) = (n-1)\left(\frac{1-a}{n}\right)^{p/(p+1)} + \left(\frac{a(1-a)(n-1)^2}{n}\right)^{p/(p+1)},$$

with the constraints $a + (1-a)/n \geq 1/(n-1)$, or equivalently $a \geq 1/(n-1)^2$. This objective is concave in a. Taking first-order conditions and simplifying, we have

$$1 - 2a^* = (a^*)^{1/(p+1)}(n-1)^{-(p-1)/(p+1)}. \tag{A18}$$

The left hand side is strictly decreasing and positive on $(0, 0.5)$ while the right-hand side is weakly increasing and always positive, so the $a^* \in (0, 0.5)$ satisfying this is unique. Moreover, this solution satisfies the feasibility constraint: at $a = 1/(n-1)^2$, the right-hand side equals $1/(n-1) \leq 1/2$ while $1 - 2/(n-1)^2 \geq 1/2$ for $n \geq 3$; thus, $a^* \geq 1/(n-1)^2$.

We now show the remaining claims. We prove $\mathbb{E}_\delta[g_k] \neq 1/2$ by contradiction. If $\mathbb{E}_\delta[g_k] = 1/2$, $\mathbb{E}_\delta[T] = n/2 > 1$ by exchangeability, while we have shown that $\mathbb{E}_\delta[T] = 1 - a^* \leq 1$.

To prove (a), note that when $n = 3$ the above first-order condition becomes

$$1 - 2a^* = (a^*)^{1/(p+1)} 2^{-(p-1)/(p+1)},$$

which is uniquely solved by $a^* = 0.25$. Similarly, note the first-order condition at $p = 1$ is

$$1 - 2a^* = \sqrt{a^*}$$

which is uniquely solved by $a^* = 0.25$. For (b), let $p_n \to \infty$ be any sequence along which $a^*_{p_n} \to a^*_\infty$; since we know

$$\begin{aligned} & 1 - 2a^*_{p_n} \le (n-1)^{-(p-1)/(p+1)} \\ \implies & a^*_{p_n} \ge \frac{1 - (n-1)^{-(p-1)/(p+1)}}{2} \to \frac{n-2}{2(n-1)} > 0, \end{aligned}$$

we have by passing through the limit:

$$1 - 2a^*_\infty = \lim_{p_n \to \infty} (a^*_{p_n})^{1/(p+1)} (n-1)^{-(p-1)/(p+1)} = (n-1)^{-1}$$

Hence $a^*_\infty = (n-2)/(2(n-1))$; (c) similarly follows taking limits of $n \to \infty$ with $p > 1$.

C.8 Proposition A11

We prove the more general result with multiple exposures: let $J(\Sigma) = \min_{c \in \mathbb{R}} \operatorname{tr}\left(((W - cU)\Sigma(W - cU)')^{p/(p+1)}\right)$ and let

$$\Sigma^* \in \arg\max_{\Sigma \in \mathcal{Q}_K} J(\Sigma),$$

where $\mathcal{Q}_K$ and $\mathcal{C}_K$ are as before; we again implement Σ^* by Gaussian rounding with Σ^{GR} denoting the variance of $g^{\text{GR}} - \frac{1}{2}\mathbf{1}$. We prove that

$$J(\Sigma^{GR}) \ge \left(\frac{2}{\pi}\right)^{p/(p+1)} \max_{\Sigma \in \mathcal{C}_K} J(\Sigma).$$

Let $R = 4\Sigma^*$ be the relaxed correlation matrix. Under Gaussian rounding, we have $\Sigma^{GR} = \frac{1}{2\pi} \arcsin(R)$ where $\arcsin(\cdot)$ is applied entrywise. By the power series

$$\arcsin(x) = x + \sum_{m=1}^{\infty} a_m x^{2m+1}, \ a_m \ge 0$$

we have entrywise:

$$\arcsin(R) - R = \sum_{m=1}^{\infty} a_m R^{\circ(2m+1)}$$

where each Hadamard power $R^{\circ(2m+1)}$ is positive semidefinite by the Schur product theorem (the Hadamard product of two positive semidefinite matrices is also positive semidefinite). Hence

$\arcsin(R) - R \succeq 0$ and

$$\Sigma_{GR} \succeq \frac{1}{2\pi} R = \frac{2}{\pi} \Sigma^*.$$

For any c, this gives:

$$(W - cU)\Sigma_{GR}(W - cU)' \succeq \frac{2}{\pi}(W - cU)\Sigma^*(W - cU)'.$$

Since $p/(p+1) \in (0, 1]$, the map $A \mapsto \mathrm{tr}(A^{p/(p+1)})$ is monotone on the positive semidefinite cone. Hence, for any fixed c:

$$\mathrm{tr}\left(((W - cU)\Sigma_{GR}(W - cU)')^{p/(p+1)}\right) \geq \left(\frac{2}{\pi}\right)^{p/(p+1)} \mathrm{tr}\left(((W - cU)\Sigma^*(W - cU)')^{p/(p+1)}\right).$$

Because this inequality holds pointwise, we have by minimizing c on both sides:

$$J(\Sigma^{GR}) \geq \left(\frac{2}{\pi}\right)^{p/(p+1)} J(\Sigma^*) \geq \left(\frac{2}{\pi}\right)^{p/(p+1)} \max_{\Sigma \in \mathcal{C}_K} J(\Sigma),$$

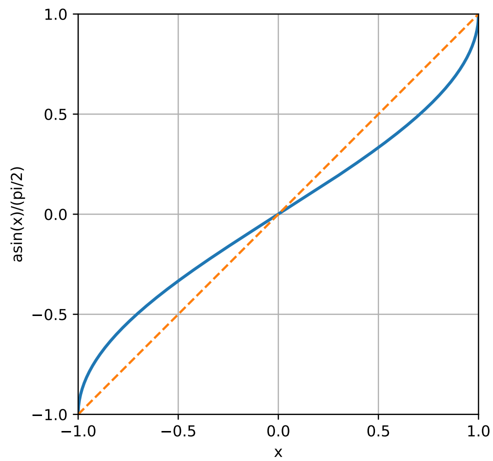
and thus

$$N^{1/p}\sigma^2 J(\Sigma^{GR})^{-(p+1)/p} \leq \frac{\pi}{2} \cdot N^{1/p}\sigma^2 \left(\max_{\Sigma \in \mathcal{C}_K} J(\Sigma)\right)^{-(p+1)/p}.$$

Here by Theorem 1 the left-hand side is the worst-case approximate variance corresponding to the Gaussian rounding design while the right-hand side is the lowest possible worst-case approximate variance.

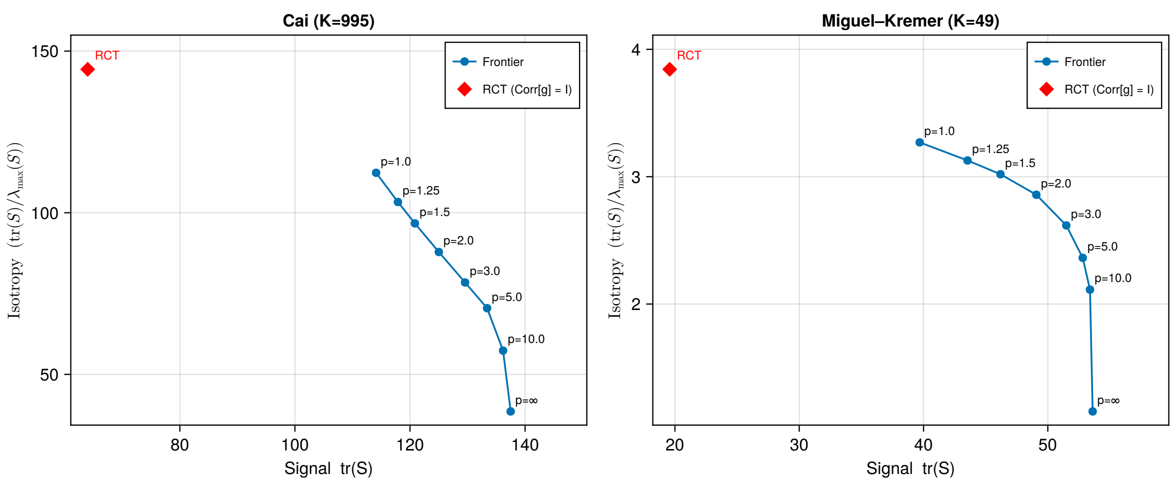
Additional Exhibits

Figure A1: Correlation Distortion from Gaussian Rounding



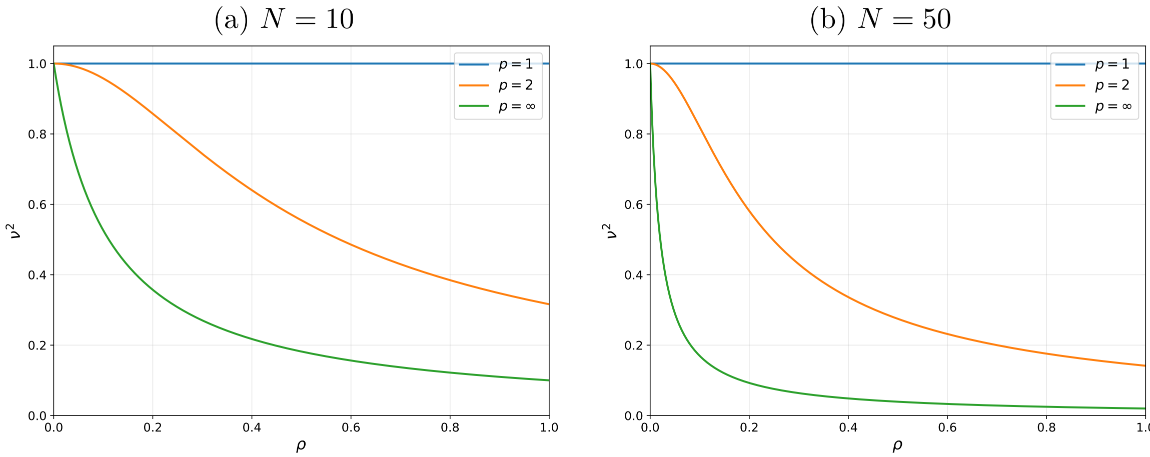
Notes: For each value of $\mathrm{Corr}[\xi_k, \xi_l]$ on the horizontal axis, the vertical axis shows $\mathrm{Corr}[g_k, g_l]$ from the Gaussian rounding procedure of Section 3.2. The dashed 45-degree line indicates the benchmark of no distortion.

Figure A2: Signal and Isotropy Frontiers



Notes: This figure shows the isotropy and signal of optimal designs, for different values of p. Signal and isotropy of the baseline of independent Bernoulli assignment is included for reference. Signal is defined as $\mathrm{tr}(S)$ for $S = \mathrm{Var}_\delta[x]$, while isotropy is defined as $\mathrm{tr}(S)/\lambda_{\max}(S)$.

Figure A3: The Tradeoff Between Variance and Mutual Correlation in the Equicorrelated Setting



Notes: In the setting of Appendix B.3, panel (a) shows the maximum variance of errors ν^2, assumed equal across all $N = 10$ observations, such that the error distribution with equal mutual correlations of errors equal to $\rho \in [0, 1]$ is within $\mathcal{F}_p(\sigma)$ for $\sigma = 1$. The computation follows condition (A14). Each line corresponds to a different choice of p. Panel (b) repeats this analysis for $N = 50$.

Table A1: RMSE and Effective Sample Size Gains for the Cai et al. (2015) Application, Without Whitening

Design and Estimator:	Homo-skedastic (1)	Hetero-skedastic (2)	Deg.-in-Mean (3)	Village-Corr. (4)	Network-Corr. (5)	Estimated Residuals (6)
RCT Benchmark	0.150	0.139	0.142	0.138	0.141	0.129
	[+0%]	[+0%]	[+0%]	[+0%]	[+0%]	[+0%]
Optimal, $p = 1$	0.116	0.119	0.124	0.118	0.113	0.102
	[+68%]	[+36%]	[+29%]	[+36%]	[+54%]	[+60%]
Optimal, $p = 2$	0.110	0.109	0.128	0.111	0.114	0.102
	[+85%]	[+61%]	[+23%]	[+55%]	[+52%]	[+60%]
Optimal, $p = \infty$	0.095	0.099	0.133	0.117	0.107	0.099
	[+151%]	[+96%]	[+13%]	[+39%]	[+73%]	[+71%]

Notes: For the Cai et al. (2015) application, this table reports the root mean-squared error of different designs (rows) under different data-generating processes for the errors (columns), using the unwhitened recentered IV estimator. The first row is the baseline of independent Bernoulli assignment while the other rows are the Algorithm 1 feasible optimal designs for different values of p (but not the optimal IV). The text describes the six error DGPs. Increases in effective sample sizes, given in brackets, are computed as the squared ratio of inverse RMSE under the focal scenario relative to the RCT baseline with the same error distributions, minus one.

Table A2: RMSE and Effective Sample Size Gains for the Miguel and Kremer (2004) Application, Without Whitening

(a) Spillover Effect

Design and Estimator:	Homo-skedastic (1)	Hetero-skedastic (2)	Deg.-in-Mean (3)	Cluster-Corr. (4)	Network-Corr. (5)	Estimated Residuals (6)
RCT Benchmark	0.101 [+0%]	0.099 [+0%]	0.128 [+0%]	0.106 [+0%]	0.109 [+0%]	0.179 [+0%]
Optimal, $p = 1$	0.070 [+112%]	0.076 [+70%]	0.072 [+214%]	0.075 [+98%]	0.074 [+118%]	0.139 [+66%]
Optimal, $p = 2$	0.064 [+148%]	0.071 [+97%]	0.064 [+293%]	0.072 [+117%]	0.071 [+135%]	0.128 [+95%]
Optimal, $p = \infty$	0.052 [+277%]	0.056 [+215%]	0.049 [+570%]	0.067 [+147%]	0.057 [+267%]	0.111 [+160%]

(b) Direct Effect

Design and Estimator:	Homo-skedastic (1)	Hetero-skedastic (2)	Deg.-in-Mean (3)	Cluster-Corr. (4)	Network-Corr. (5)	Estimated Residuals (6)
RCT Benchmark	0.052 [+0%]	0.049 [+0%]	0.040 [+0%]	0.041 [+0%]	0.034 [+0%]	0.064 [+0%]
Optimal, $p = 1$	0.053 [-2%]	0.051 [-10%]	0.041 [-4%]	0.046 [-20%]	0.035 [-7%]	0.066 [-6%]
Optimal, $p = 2$	0.053 [-4%]	0.049 [-1%]	0.038 [+11%]	0.046 [-20%]	0.039 [-24%]	0.065 [-3%]
Optimal, $p = \infty$	0.051 [+6%]	0.048 [+5%]	0.040 [+1%]	0.046 [-20%]	0.038 [-23%]	0.100 [-59%]

Notes: For the Miguel and Kremer (2004) application, panel (a) reports the root mean-squared error of different designs (rows) for the spillover effect under different data-generating processes for the errors (columns), using the unwhitened recentered IV estimator. Panel (b) reports the same for the direct effect. The first row is the baseline of independent Bernoulli assignment while the other rows are the Algorithm 1 feasible optimal designs for different values of p (but not the optimal IV). The text describes the six error DGPs. Increases in effective sample sizes, given in brackets, are computed as the squared ratio of inverse RMSE under the focal scenario relative to the RCT baseline with the same error distributions, minus one.

Table A3: Coverage for Cai et al. (2015) Simulation

Design and Estimator:	Homo-skedastic (1)	Hetero-skedastic (2)	Deg.-in-Mean (3)	Village-Corr. (4)	Network-Corr. (5)	Estimated Residuals (6)
RCT Benchmark	0.950	0.949	0.947	0.949	0.948	0.947
$p = 1$	0.948	0.946	0.952	0.947	0.948	0.950
$p = 2$	0.948	0.950	0.950	0.951	0.947	0.949
$p = \infty$	0.948	0.945	0.946	0.948	0.947	0.950

Notes: For the Cai et al. (2015) application, this table reports the empirical coverage for the whitened recentered IV estimator with different designs (rows) and different data-generating processes for the errors (columns), using a confidence interval based on the normal approximation $\hat{\beta} \pm 1.96 \cdot \frac{\sqrt{\hat{v}}}{\sqrt{K}|h_K|}$, as defined in Section 3.3. Coverage is computed as the fraction of times in 10,000 replications of the Monte-Carlo simulation that the confidence interval contains the true β.

Table A4: Coverage for the Miguel and Kremer (2004) Application

(a) Spillover Effect

Design and Estimator:	Homo-skedastic (1)	Hetero-skedastic (2)	Deg.-in-Mean (3)	Cluster-Corr. (4)	Network-Corr. (5)	Estimated Residuals (6)
RCT Benchmark	0.966	0.965	0.935	0.949	0.941	0.935
$p = 1$	0.964	0.960	0.937	0.955	0.947	0.936
$p = 2$	0.952	0.950	0.932	0.933	0.926	0.918
$p = \infty$	0.651	0.655	0.609	0.491	0.590	0.424

(b) Direct Effect

Design and Estimator:	Homo-skedastic (1)	Hetero-skedastic (2)	Deg.-in-Mean (3)	Cluster-Corr. (4)	Network-Corr. (5)	Estimated Residuals (6)
RCT Benchmark	0.953	0.951	0.958	0.951	0.953	0.951
$p = 1$	0.947	0.951	0.955	0.945	0.945	0.938
$p = 2$	0.946	0.943	0.951	0.942	0.941	0.932
$p = \infty$	0.690	0.731	0.655	0.653	0.728	0.413

Notes: For the Miguel and Kremer (2004) application, panel (a) reports the empirical coverage for the whitened recentered IV estimator for the indirect effect with different designs (rows) and different data-generating processes for the errors (columns), using a confidence interval based on the normal approximation, based on an extension of Theorem 2 to multiple exposures. Panel (b) reports the same for the direct effect. For each panel, coverage is computed as the fraction of times in 10,000 replications of the Monte-Carlo simulation that the confidence interval contains the true target parameter.