# IMPROVED GUARANTEES FOR HETEROGENEOUS TREATMENT-EFFECT ESTIMATION VIA MATRIX COMPLETION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A central goal of modern causal inference is estimating heterogeneous treatment effects to answer questions like "how does an intervention affect each unit," rather than only on average. We study this problem with panel-data where we observe $n$ units across $m$ times under unknown, non-uniform treatment assignments. The data in this setting is naturally represented as a matrix of all unit–time treatment effects. Estimating heterogeneous treatment effects can then be expressed as obtaining a good estimation of each row's average in this matrix. This allows us to formulate the problem as matrix completion, which can be solved under natural low-rankness assumptions. However, existing matrix-completion guarantees are not powerful enough to get meaningful bounds for the per-row guarantee required for estimating the heterogeneous treatment effect; roughly speaking, they are only useful for estimating *average* treatment effect bounds, as also illustrated in a recent line of work. We give a simple, computationally efficient estimator that, without knowledge of the propensities and under standard low-rankness and regularity assumptions, achieves a row-wise $\ell_2$ error of $\widetilde{O}(\sqrt{1/n + n/(m^2)})$. Technically, our analysis establishes the first sharp row-wise $\ell_2$-perturbation bound for low-rank approximation, complementing existing spectral-, Frobenius-, and entrywise perturbation theory. We evaluate our estimator on both synthetic and semi-synthetic data, and find lower estimation error than direct Horvitz–Thompson and Hájek estimators for several unit-specific treatment effects. Further, in synthetic experiments, the error decay matches the dependence on panel size predicted by our theoretical bounds.

## 1 INTRODUCTION

A central question in modern causal inference is to understand for whom and when a treatment works, not just whether it works on average (Athey & Imbens, 2016). As Kravitz et al. (2004) put it, *"the benefit or harm of [average] treatments in clinical trials can be misleading and fail to reveal the potentially complex mixture of substantial benefits for some, little benefit for many, and harm for a few."* Estimating such heterogeneous effects requires data that observes the same units (representing, *e.g.*, users, patients, customers, or regions) across many randomized occasions, and a range of modern experimental designs produces data of precisely this form, including mobile-health trials, sequential A/B tests, contextual bandit experiments, and switchback experiments (Johari et al., 2022; Li et al., 2010; Bojinov et al., 2023). In all of these settings, we observe a panel of $n$ units across $m$ times, where $m$ is typically larger than $n$ (*i.e.*, $m \geq n$). Each unit–time pair $(i, j)$ has two potential outcomes: $Y_{ij}(1)$ which arises under treatment and $Y_{ij}(0)$ which arises in the absence of treatment, *i.e.*, in control. We observe only one outcome, the one corresponding to the assignment that was actually made (Imbens & Rubin, 2015). Stacking across units and times yields two outcome matrices $Y(1), Y(0) \in \mathbb{R}^{n\times m}$ in which exactly one entry from each pair $(Y_{ij}(1), Y_{ij}(0))$ is observed. The other entry is missing and is the counterfactual we would like to estimate.

We focus on the randomized setting, in which the treatment indicator $D_{ij} \in \{0, 1\}$ is an independent Bernoulli draw with probability $p_{ij} \in (0, 1)$. These probabilities can vary across units and times, reflecting different protocols, contexts, or eligibility rules. Further, they are, in general, unknown to the analyst: this is because, in many logged or retrospective experiments the assignment matrix is recorded but the platform's specific traffic splits, randomization rules, and ramp-up schedules are



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

not (Schnabel et al., 2016; Swaminathan & Joachims, 2015). Canonical examples are mobile-health trials that repeatedly randomize each participant, known as *micro-randomized trials* (Klasnja et al., 2015; Liao et al., 2016; Boruvka et al., 2018; Qian et al., 2022).

As a concrete example, consider the HeartSteps study (Klasnja et al., 2019). Here, researchers followed $n = 44$ participants for six weeks, randomizing each at up to five points per day among three options (no notification, a walking suggestion, or a stand-up suggestion) and recorded the participant's step count over the following thirty minutes as the outcome; this yielded $m = 210$ times per participant. The average treatment effect, aggregated across all participants and notification types, was not statistically significant. However, a finer analysis revealed striking heterogeneity: for instance, the lift from walking suggestions, which initially more than doubled step count, decayed substantially over the six weeks, while stand-up suggestions had no detectable effect throughout. Such findings guide interventions, such as reducing the frequency of walking suggestions as a trial progresses, and are visible only through estimates of *heterogeneous* treatment effects.

**Signal–Noise Model.** A substantial line of work in causal inference is devoted to estimating heterogeneous treatment effects (Athey & Imbens, 2016; Wager & Athey, 2018; Chernozhukov et al., 2018; Künzel et al., 2019). To formalize the treatment effect, we adopt the standard signal-plus-noise model $Y(a) = A(a) + E(a)$ for $a \in \{0, 1\}$, where $\mathbb{E}[E(a)] = 0$, so that $A(a) \coloneqq \mathbb{E}[Y(a)]$ is the expected potential outcome under action $a$ for each unit at each time. The expectation is taken over noise from the environment and other unobserved factors. The natural object for heterogeneous-effect estimation is then the unit–time treatment-effect matrix

$$M \coloneqq A(1) - A(0) = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]\,,$$

whose row $M_i = (M_{i1}, M_{i2}, \ldots, M_{im})$ records how unit $i$'s response to the treatment varies over time. For each unit $i$, one can define several heterogeneous treatment effects by computing averages $\mathrm{Avg}_i(S) \coloneqq \sum_{j \in S} M_{ij} / |S|$ over subsets $S \subseteq [m]$. For instance, if $S = [m]$, then $\mathrm{Avg}_i(S)$ is the treatment effect specifically for unit $i$ averaged over all times. One can also select $S$ to be other meaningful subsets such as the first week of a trial, weekends, or high-engagement days.

**Matrix completion for causal inference.** Estimating the matrix $M$ naturally decomposes into two estimation problems, one for $\mathbb{E}[Y(1)]$ and one for $\mathbb{E}[Y(0)]$. Fix any $a \in \{0, 1\}$: estimating $\mathbb{E}[Y(a)]$ is a matrix-completion problem, where the goal is to recover the missing entries from those observed. Without further structure on the signal matrices $A(0)$ and $A(1)$ this task is impossible; however, real-world matrices have natural structural constraints that make the task tractable. In particular, Athey et al. (2021) proposed utilizing (approximate) low-rankness of the potential output matrices. Intuitively, low-rankness amounts to assuming that unit responses are governed by a small number of latent factors (such as baseline activity level, receptiveness to prompts, time-of-day patterns, habituation, or seasonality). This assumption is well grounded in the causal panel data literature, where a long line of work builds on precisely this latent-factor structure (Abadie & Gardeazabal, 2003; Abadie et al., 2010; Xu, 2017; Arkhangelsky et al., 2021; Bai & Ng, 2021). Casting the problem as matrix completion has the further benefit of bringing to bear a broad and well-developed algorithmic toolkit, *e.g.*, (Candès & Recht, 2009; Candès & Tao, 2010; Mazumder et al., 2010).

**Limitations of prior guarantees.** Following Athey et al. (2021), a growing line of work has applied matrix completion to treatment effect estimation under low-rankness and other standard regularity requirements (Xu, 2017; Bai & Ng, 2021; Arkhangelsky et al., 2021; Agarwal et al., 2023). The guarantees produced by these works, however, are not strong enough for our purpose:

- They either control the error in Frobenius norm, $\|\widehat{M} - M\|_{\mathrm{F}} / \sqrt{nm}$, as in Athey et al. (2021). This is insufficient to control the above heterogeneous effects $\mathrm{Avg}_i(S)$ because, *e.g.*, a single unit $i$ with constant row-normalized error $\|\widehat{M}_i - M_i\|_2 / \sqrt{m} = \Omega(1)$ contributes only $O(1/\sqrt{n})$ to the normalized Frobenius error, and so leaves $\mathrm{Avg}_i(S)$ uncontrolled at that unit.
- Or they control the worst-case entrywise error $\max_{i,j} |\widehat{M}_{ij} - M_{ij}|$, as in Agarwal et al. (2023). While this is strong enough to bound $\mathrm{Avg}_i(S)$, the best known bounds on this entrywise error decay only as $1/\sqrt{\log m} + 1/\sqrt{\log n}$ (Agarwal et al., 2023), which is prohibitively slow.[1] Indeed, achieving an error of, *e.g.*, 0.1 requires an astronomically large $n, m \gtrsim e^{100}$.

[1] They incur this slow rate because their rate depends on the size of the largest all-1s sub-matrix of $D$ which, in our setting, has sides $O(\sqrt{\log n})$ and $O(\sqrt{\log m})$ with high probability.

Thus, in our setting, none of the existing guarantees yields a meaningful bound. Let $\widehat{M}$ be an estimator of $M$. A natural way to estimate $\mathrm{Avg}_i(S)$ for each unit $i$ and subsets $S$ of interest is to satisfy:

$$(1/\sqrt{m}) \cdot \|M - \widehat{M}\|_{2,\infty} := (1/\sqrt{m}) \cdot \max_{1\le i\le n} \|M_i - \widehat{M}_i\|_2 \le \mathrm{poly}(1/n, 1/m)\,. \tag{1.1}$$

This immediately gives bounds on the estimation error of $\mathrm{Avg}_i(S)$. Let $\widehat{\mathrm{Avg}}_i(S) := \sum_{j\in S} \widehat{M}_{ij}/|S|$. Now Cauchy–Schwarz inequality shows that, for every unit $i$ and subset $S \subseteq [m]$,

$$|\widehat{\mathrm{Avg}}_i(S) - \mathrm{Avg}_i(S)| \;\le\; \sqrt{1/|S|} \cdot \|M - \widehat{M}\|_{2,\infty} \;\le\; \sqrt{m/|S|} \cdot \mathrm{poly}(1/n, 1/m)\,.$$

Hence, for any unit $i$ and every subset $S$ with $|S| = \Omega(m)$, $|\widehat{\mathrm{Avg}}_i(S) - \mathrm{Avg}_i(S)| \le \mathrm{poly}(1/n, 1/m)$.

This leads us to the central question in the paper: Can such a row-wise guarantee be achieved from the observed entries, under existing structural assumptions? A bit more formally:

> **Question.** *Given the observed outcomes, can one efficiently compute a matrix $\widehat{M}$ that, under existing low-rankness and regularity conditions, satisfies $(1/\sqrt{m})\|M - \widehat{M}\|_{2,\infty} \le \mathrm{poly}(1/n, 1/m)$?*

### 1.1 Our contributions

We answer the above question affirmatively, providing an estimator that achieves $(1/\sqrt{m})\|M - \widehat{M}\|_{2,\infty} \lesssim \sqrt{1/n + 1/m}$ with high probability under standard low-rankness and regularity conditions. Concretely, we work under the following regularity conditions on the signal $A$ and noise $E$.

**Informal Assumption 1.1** (Signal and noise regularity)**.** *For each action $a \in \{0,1\}$, let $\sigma_1(a) \ge \sigma_2(a) \ge \dots$ denote the singular values of $A(a)$, and let $U(a), V(a)$ collect the left and right singular vectors corresponding to its leading $r$ singular components. The following hold for each $a \in \{0,1\}$:*

- ***(Approximate low rankness)*** *$A(a)$ has at most $r$ "large" singular values: $\sigma_{r+1}(a) \lesssim K\sqrt{m+n}$.*
- ***(Bounded signal and noise)*** *The signal and noise are bounded: $\|A(a)\|_\infty, \|E(a)\|_\infty \le K$.*
- ***(Independent and mean-zero noise)*** *The entries of $E(a)$ are independent and mean-zero.*
- ***(Row and column incoherence)*** *The leading left and right singular spaces are "incoherent:"*
$$\sqrt{n}\max_{1\le i\le n} \|e_i^\top U(a)\|_2 \le \mu \quad \text{and} \quad \sqrt{m}\max_{1\le j\le m} \|e_j^\top V(a)\|_2 \le \mu\,.$$
- ***(Signal-to-noise ratio)*** *The first singular value of $A$ is "large:" $\sigma_1(a) \gtrsim rK\sqrt{n+m}$.*

As mentioned, approximate low-rankness is standard and well motivated Athey et al. (2021). Further, boundedness and mean-zero noise are mild regularity requirements satisfied in most practical settings. The incoherence and signal-to-noise assumptions are regularity assumptions that can also shown to be necessary. Row incoherence rules out the degenerate case in which the signal is concentrated on a few units; and, due to this, the information from these units' rows is not reliably available from the rest of the rows. Column incoherence is the analogous requirement across time. Finally, the lower bound on the signal-to-noise ratio is mild and needed for the algorithm to reliably estimate the rank $r$ from the observed data; it is widely used in matrix perturbation work (Stewart & Sun, 1990).

We need some additional notation to state our result. For each action $a \in \{0,1\}$, define $p_{ij}(a) := \Pr(D_{ij} = a)$, so that $p_{ij}(1) = p_{ij}$ and $p_{ij}(0) = 1 - p_{ij}$. Let $\overline{p}_i(a) := (1/m)\sum_{j=1}^m p_{ij}(a)$ be the average probability of observing $Y(a)$ for unit $i$. We define

$$q := \min_{a,i}\ \overline{p}_i(a), \quad r_p := \max_{a,i,j}\ \frac{p_{ij}(a)}{\overline{p}_i(a)}, \quad \text{and} \quad P(a)_{ij} := \frac{p_{ij}(a)}{\overline{p}_i(a)} - 1\,. \tag{1.2}$$

Here, $q$ is the smallest average observation rate for any unit $i$ and action $a$. The parameter $r_p$ measures how uneven the observation probabilities can be within a row after normalizing by the row average. Finally, $P(a)$ measures the remaining within-row nonuniformity of the design. To gain some intuition, consider the special case where all observation probabilities are constant, $p_{ij} = c$. In this case, $q = c$, $r_p = 1$, and $P(a) = 0$ for each $a$.

**Theorem 1.2** (Main guarantee, informal)**.** *Under Informal Assumption 1.1, there is a polynomial-time algorithm which, given only the observed outcomes $Y_{\mathrm{obs}}$ and the assignment matrix $D$, and* without any knowledge of the propensities $p_{ij}$*, outputs an estimate $\widetilde{M}$ satisfying, with high probability,*

$$\frac{1}{\sqrt{m}} \cdot \|M - \widehat{M}\|_{2,\infty} \;\le\; \widetilde{O}\left( Kr^{3/2}\mu \left[ \sqrt{\frac{r_p}{q}\left(\frac{1}{n} + \frac{n}{m^2}\right)} + \frac{\max_a \|P(a)\|_{\mathrm{op}}}{\sqrt{m \cdot \min\{m,n\}}} \right] \right)\,.$$

In the simplest case where $p_{ij} = c$ (for each $i, j$) and $K, r, \mu = O(1)$, the main theorem yields

$$\frac{1}{\sqrt{m}} \cdot \|M - \widehat{M}\|_{2,\infty} \leq \widetilde{O}\left(\sqrt{\frac{1}{n} + \frac{n}{m^2}}\right),$$

which achieves the desired $\mathrm{poly}(1/n, 1/m)$ rate in the standard regime where $m \gtrsim n$. In contrast, in the same setting the prior works either do not provide any non-trivial bound (Athey et al., 2021) or only achieve a $\mathrm{poly}(1/\log n, 1/\log m)$ rate Agarwal et al. (2023). The same rate holds for the *row-homogeneous design*, where $p_{ij} = p_i$ for all $j$, *i.e.*, different units may be randomized at different rates, but each unit's rate is stable over time (Corollary 3.4).

**Technical Novelty (also see Section 4).** One important advantage of the estimator that we propose is that it is very simple and easy to implement: is a simple row-scaled truncated-SVD algorithm (Algorithm 1). The analysis of this simple algorithm in our setting though, requires several new tools in the spirit of modern results from matrix perturbation theory. Existing work on perturbation bounds of low-rank approximations has primarily focused on either bounds for different norms *e.g.*, Tran & Vu (2025c); Tran et al. (2025); Eckart & Young (1936); Tran & Vu (2025a); Chatterjee (2015); Mangoubi & Vishnoi (2025); Dwork et al. (2014); Bhardwaj & Vu (2024), which are not applicable in our setting, or bounds that do not gracefully improve for low-rank matrices (Cape et al., 2019; Abbe et al., 2020), and hence, are not useful for analyzing our algorithm. *Our main technical contribution is a perturbation theory for truncated SVD in the $\|\cdot\|_{2,\infty}$ norm, which directly controls the reconstructed error of our algorithm.* To obtain the required sharp bound, we need to carefully adapt the contour expansion method, introduced in Tran & Vu (2026a; 2025a); Tran et al. (2025) and applied in the more challenging norm $\|\cdot\|_{2,\infty}$. In particular, the contour bootstrapping argument in Tran & Vu (2026a; 2025a); Tran et al. (2025) does not apply here, so we need a new idea to control the contour expansion error. See Section 4 for details.

A second difficulty comes from the fact that the probabilities $p_{ij}$ are unknown and nonuniform. If the entrywise probabilities $p_{ij}(a)$ were known, one could form an unbiased inverse-propensity-weighted matrix by scaling each observed entry by $1/p_{ij}(a)$. Our algorithm does not have this information and instead scales row $i$ by its empirical observation frequency. The population analogue of this row scaling is unbiased when the design is row-homogeneous (*i.e.*, $p_{ij} = p_{ik}$ for each $j, k$), but under nonuniform propensities its expectation satisfies $p_{ij}(a)A_{ij}(a)/\overline{p}_i(a) = A_{ij}(a) + P_{ij}(a)A_{ij}(a)$. Thus the perturbation is not purely mean-zero noise: it also contains a deterministic bias induced by nonuniformity within one row. Our analysis incorporates this bias directly into the perturbation argument, yielding a bound whose additional design-dependent term is governed by $\|P(a)\|_{\mathrm{op}} / \sqrt{m \cdot \min(m, n)}$, which decays as $m, n \to \infty$ and vanishes when $P(a) = 0$.

### 1.2 Additional Related Works

Our work is broadly related to two lines of work: (1) work utilizing low-rank approximation results in causal inference and (2) work on matrix perturbation bounds for low-rank approximations. We have already mentioned key works from both lines in the introduction. Here we give a more detailed comparison with the first line. A more technical comparison with the second line appears in Technical Overview (Section 4). A growing literature uses low-rank structure to estimate missing counterfactual outcomes in panel data. As we have already mentioned, closest to our work are Athey et al. (2021); Agarwal et al. (2023), and their guarantees are insufficient for the heterogeneous effects we study. Several other works also exploit low-rankness in panel outcomes but their goal is to estimate (different types of) *averages* treatment (Amjad et al., 2018; 2019; Arkhangelsky et al., 2021; Bai & Ng, 2021; Fernández-Val et al., 2021). In contrast, we estimate heterogeneous treatment effects. A different line of works imposes low-rankness on the covariate matrix (rather than on the potential-outcome matrices) and use it to impute or de-noise covariates (Agarwal et al., 2021; 2025a), while they also use low-rankness, their estimation targets and setting is quite different from our work. Finally, Agarwal et al. (2025b) extend the low-rank framework to settings with more than two treatment conditions and give row-wise guarantees for estimating $\mathrm{Avg}_i([m])$. Our row-wise control of $M$ is stronger: it yields bounds on $\mathrm{Avg}_i(S)$ for every sufficiently large $S \subseteq [m]$, not just $S = [m]$.

## 2 Preliminaries

**Notation.** For constants $N$ and $1 \leq k \leq N$, let $e_{N,k}$ be the $k$th standard basis vector in $\mathbb{R}^N$; when the dimension is clear, we write $e_k$. For a vector $v$ and $p \geq 1$, define $\|v\|_p$ as $(\sum_j |v_j|^p)^{1/p}$ and

$\|v\|_\infty$ as $\max_j |v_j|$. For a matrix $B \in \mathbb{R}^{n\times m}$, we write $B_i$ for its $i$th row, $B_{\cdot j}$ for its $j$th column, $\|B\|_\infty$ for $\max_{i,j}|B_{ij}|$, and $\|B\|_{\mathrm{op}}$ and $\|B\|_{\mathrm{F}}$ for its operator and Frobenius norms respectively. The central object for our row-wise guarantees is the row-wise $\ell_2$-norm $\|B\|_{2,\infty} := \max_{1\le i\le n}\|B_i\|_2$.

**Potential Outcomes and the Heterogeneous Treatment Effect.** Following the potential-outcomes framework (Imbens & Rubin, 2015; Hernan & Robins, 2023), $Y(a) = A(a) + E(a)$ with $\mathbb{E}[E(a)] = 0$, and our goal is to estimate $M := A(1) - A(0)$. Unless otherwise stated, expectations are taken over the outcome noise $E$, not the treatment assignments.

**Assignment and Observed Data.** We make standard assumptions on the assignment mechanism Hernan & Robins (2023); Imbens & Rubin (2015). For each $(i, j)$, the treatment indicator $D_{ij} \in \{0, 1\}$ is an independent Bernoulli draw, $D_{ij} \sim \mathrm{Ber}(p_{ij})$, where the probabilities $p_{ij}$ may vary across units and times and are not assumed to be known. The assignments $D = (D_{ij})$ are independent of the potential outcomes $\{Y(0), Y(1)\}$, and the observed outcome at $(i, j)$ is $Y^{\mathrm{obs}}_{ij} := D_{ij}Y_{ij}(1) + (1 - D_{ij})Y_{ij}(0)$. To treat the two actions symmetrically, we introduce,

$$D_{ij}(1) := D_{ij}\,, \quad D_{ij}(0) := 1 - D_{ij}\,, \quad p_{ij}(1) := p_{ij}\,, \quad p_{ij}(0) := 1 - p_{ij}\,,$$

so that $D_{ij}(a) = 1$ when $Y_{ij}(a)$ is observed and $\mathbb{E}[D_{ij}(a)] = p_{ij}(a)$. The corresponding partially observed matrix $\widetilde{Y}(a)$ is defined entrywise by

$$\widetilde{Y}_{ij}(a) := Y_{ij}(a) \quad \text{if} \quad D_{ij}(a) = 1 \qquad \text{and} \qquad \widetilde{Y}_{ij}(a) := \star \quad \text{otherwise}\,,$$

with $\star$ denoting a missing entry. The analyst equivalently sees the pair $(D, Y^{\mathrm{obs}})$, or the pair of partially observed matrices $(\widetilde{Y}(0), \widetilde{Y}(1))$ together with their observation masks.

**Row Propensities and Non-uniformity.** The rates we obtain depend on the parameters $q$, $r_p$, and $P(a)$ defined in Equation (1.2). The parameter $q$ controls how often the rarest action is observed at the rarest unit. Requiring $q \ge \Omega(1)$ is substantially milder than the standard overlap condition (Imbens & Rubin, 2015; Hernan & Robins, 2023), which requires every $p_{ij}$ to be bounded away from 0 and 1; our condition instead constrains only the row averages, leaving individual $p_{ij}$ free to be arbitrarily close to 0 or 1 at many time points within a unit. The parameters $r_p$ and $P(a)$ quantify within-row nonuniformity. By construction, each row of $P(a)$ averages to zero, and $P(0) = P(1) = 0$ exactly when the design is row-homogeneous, *i.e.*, $p_{ij} = p_i$ for all $i, j$.

**SVD, Low-Rankness, and Noise.** For a rank-$r$ matrix $A \in \mathbb{R}^{n\times m}$, write $A = U\Sigma V^\top = \sum_{\ell=1}^{r}\sigma_\ell u_\ell v_\ell^\top$, where $U = [u_1, \dots, u_r] \in \mathbb{R}^{n\times r}$ and $V = [v_1, \dots, v_r] \in \mathbb{R}^{m\times r}$ have orthonormal columns and the singular values $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r$ are arranged in non-increasing order. We write $A_s$ for the best rank-$s$ approximation of $A$, and

$$\delta_s(A) := \sigma_s - \sigma_{s+1}$$

for its $s$th singular value gap; when $A$ is clear from context we abbreviate $\delta_s := \delta_s(A)$. The row- and column-incoherence parameters of $A$ are

$$\mu_R(A) := \sqrt{n}\max_{1\le i\le n}\|e_{n,i}^\top U\|_2 \qquad \text{and} \qquad \mu_C(A) := \sqrt{m}\max_{1\le j\le m}\|e_{m,j}^\top V\|_2\,, \tag{2.1}$$

and we set $\mu(A) := \max\{\mu_R(A), \mu_C(A)\}$. These quantities are small when the singular spaces spread across units and time points, and large when either side concentrates on few coordinates. When both potential-outcome signals are under consideration we write $\mu := \max_{a\in\{0,1\}}\mu(A(a))$.

**Definition 2.1** ($(K, \sigma)$-Bounded Random Matrix)**.** A random matrix $E_R \in \mathbb{R}^{n\times m}$ is called $(K, \sigma)$-bounded if its entries satisfy $\mathbb{E}\,E_{R,ij} = 0$, $\mathbb{E}[|E_{R,ij}|^2] \le \sigma^2$, and $\mathbb{E}[|E_{R,ij}|^\ell] \le K^{\ell-2}\sigma^2$ for every $\ell \ge 2$ and all $i \in [n]$, $j \in [m]$.

This condition fixes the variance scale and supplies the higher-moment control used in Bernstein-type matrix concentration (Tropp, 2012; 2015).

## 3 Our Main Results

In this section, we present our main results. We begin with the formal version of Informal Assumption 1.1, followed by our estimator, and then its guarantees. Recall the parameters $q, r_p$, and $P(a)$ from Equation (1.2). For each $a \in \{0, 1\}$, write

$$T(a) := \sqrt{\frac{m + r_p n}{q}\,\log(m + n)} + \|P(a)\|_{\mathrm{op}}\,.$$

We impose the following standard regularity conditions

**Assumption 3.1** (Regularity conditions)**.** Fix constants $K_A, K_E, \mu > 0$ and $r \in \mathbb{N}$. Set $K := K_A + K_E$. For each $a \in \{0,1\}$, the following conditions hold:

1. **(Low-Rankness)** $\|A(a) - A_r(a)\|_{\text{op}} \lesssim K\sqrt{m+n}$, with $A_r(a)$ the best rank-$r$ approximation
2. **(Bounded Signal and Noise)** $\|A(a)\|_\infty \le K_A$ and $E$ is $(K_E, K_E)$-bounded (Definition 2.1).
3. **(Independent and mean-zero)** $E(a)$ has independent, mean-zero entries, independent of $D$
4. **(Row and column incoherence)** $\mu_C(A(a)), \mu_R(A(a)) \le \mu$, where $\mu(\cdot)$ is defined in (2.1).
5. **(Signal-to-noise ratio)** The leading singular value satisfies $\sigma_1(a) \gtrsim KrT(a)$.

We refer the reader to Section 1.1 and Section B.1 for discussion of this assumption.

Next, we present our estimator (see Algorithm 1). It computes $\widehat{A}(0)$ and $\widehat{A}(1)$ separately and returns $\widehat{M} := \widehat{A}(1) - \widehat{A}(0)$. To estimate $A(a)$, we fill unobserved entries with zeros, rescale each row by its empirical observation frequency, and keep the largest singular block separated by a spectral gap.

**Algorithm 1** Row-scaled spectral estimator

**Input:** Observed outcomes $Y^{\text{obs}}$, matrix $D$, $r$ from Assumption 3.1, and thresholds $\tau_0, \tau_1$.
**for each** $a \in \{0,1\}$ **do**
  Initialize the estimates $\widehat{p}_i(a) \leftarrow \max\{m^{-1}\sum_{j=1}^m D_{ij}(a),\, m^{-1}\}$ for all $i \in [n]$.
  For each $(i,j) \in [n] \times [m]$, set $Z_{ij}(a) \leftarrow D_{ij}(a)Y^{\text{obs}}_{ij}$ and $\widetilde{Y}^{\text{ub}}_{ij}(a) \leftarrow Z_{ij}(a)/\widehat{p}_i(a)$.
  Let $\widetilde{\sigma}^{\text{ub}}_1(a) \ge \widetilde{\sigma}^{\text{ub}}_2(a) \ge \dots$ be the singular values of $\widetilde{Y}^{\text{ub}}(a)$.
  Let $\mathcal{S}_a \leftarrow \{1 \le s \le r : \widetilde{\sigma}^{\text{ub}}_s(a) - \widetilde{\sigma}^{\text{ub}}_{s+1}(a) \ge \tau_a\}$.
  Set $\widehat{s}(a) \leftarrow \max \mathcal{S}_a$ if $\mathcal{S}_a \neq \emptyset$, and $\widehat{s}(a) \leftarrow 0$ otherwise.
  Set $\widehat{A}(a) \leftarrow [\widetilde{Y}^{\text{ub}}(a)]_{\widehat{s}(a)}$, where $[B]_s$ denotes the best rank-$s$ approximation of $B$
**Return** $\widehat{M} \leftarrow \widehat{A}(1) - \widehat{A}(0)$.

**Running time of Algorithm 1.** The dominant cost is computing the top $r+1$ singular values and vectors of $\widetilde{Y}^{\text{ub}}(a)$ for each $a \in \{0,1\}$, which suffice to determine $\widehat{s}(a)$ and $[\widetilde{Y}^{\text{ub}}(a)]_{\widehat{s}(a)}$. This takes $\widetilde{O}(nmr)$ time via Lanczos's method or randomized SVD (Golub & Van Loan, 2013). All remaining steps run in $O(nm + nr)$ time, giving a total running time of $\widetilde{O}(nmr)$.

**Main Result.** Next, we state our main result, which bounds the error in estimation of $A(0)$ and $A(1)$; combining the two bounds immediately implies a bound on the estimation error for $M$.

**Theorem 3.2** (Upper bound on Error)**.** *Suppose Assumption 3.1 holds. Run Algorithm 1 with thresholds $\tau_a = 96KT(a)$ for $a \in \{0,1\}$. Then, with probability at least $1 - O(1/(m+n))$, the output matrices $\widehat{A}(0)$ and $\widehat{A}(1)$ satisfy, for each $a \in \{0,1\}$,*

$$\|\widehat{A}(a) - A(a)\|_{2,\infty} \;\lesssim\; Kr^{3/2}\mu\;(\sqrt{m+n})\log^4(m+n)\;\left[\sqrt{\frac{r_p}{mq} + \frac{r_p}{nq}} + \frac{\|P(a)\|_{\text{op}}}{\sqrt{mn}}\right]. \quad (3.1)$$

Now, applying Theorem 3.2 to both $a = 0, 1$ yields the following row-wise guarantee for $\widehat{M}$.

**Corollary 3.3** (Row-wise recovery of the treatment-effect matrix)**.** *Under the assumptions of Theorem 3.2, with probability at least $1 - O(1/(m+n))$,*

$$\max_{1\le i\le n} \frac{\|\widehat{M}_i - M_i\|_2}{\sqrt{m}} \;\lesssim\; Kr^{3/2}\mu\;\log^4(m+n)\;\left[\sqrt{\frac{r_p}{q}\left(\frac{1}{n} + \frac{n}{m^2}\right)} + \frac{\max_{a\in\{0,1\}}\|P(a)\|_{\text{op}}}{\sqrt{m\min\{m,n\}}}\right].$$

Thus, when $m \ge n$, $K, r, \mu, r_p$ and $q^{-1}$ are constants, and the design is row-homogeneous (so $P(0) = P(1) = 0$), the error scales as $\widetilde{O}(n^{-1/2})$. Next, to build some intuition, we highlight the bounds we obtain under two regimes that are common in panel-data applications.

**Special Case I (Row-homogeneous design):** Our first special case is:

$$p_{ij} = p_i \qquad \text{for every } i \in [n] \text{ and } j \in [m]. \quad (3.2)$$

This setting captures stratified and covariate-adaptive randomization protocols, and also includes the Bernoulli design $p_{ij} = c$ as a sub-case. It is useful because under any row-homogeneous design, $P(0) = P(1) = 0$ and $r_p = 1$, leading to the following bound.

**Corollary 3.4** (Row-homogeneous design)**.** *Consider the row-homogeneous design above (Equation (3.2)). Suppose the assumptions of Theorem 3.2 hold, $K, r, \mu = O(1)$, $q = \Omega(1)$, and $m \geq n$. Then, with probability $1 - O(1/(m+n))$ it holds that $(1/\sqrt{m}) \cdot \|\widehat{M} - M\|_{2,\infty} \leq \widetilde{O}(1/\sqrt{n})$.*

In the balanced regime $m \asymp n$, this matches the best known guarantees for matrix completion under the Bernoulli design, despite the fact that our algorithm does not have access to the propensities $p_i$.

**Special Case II (Spectrally small within-row non-uniformity):** Our next regime is more general: it allows $p_{ij}$ to vary in time within rows but requires the variation to be *spectrally* small. Concretely,

$$\text{for some } \nu \geq 0\,, \quad \max_{a\in\{0,1\}} \|P(a)\|_{\mathrm{op}} \leq \nu\, \widetilde{O}(\sqrt{m} + \sqrt{n})\,. \tag{3.3}$$

For example, this holds whenever $P(a)$'s entries are independent and sub-Gaussian, as in contextual-bandit-style adaptive designs in which propensities depend sufficiently mildly on observed covariates; standard random-matrix bounds then give the spectral condition with high probability.

**Corollary 3.5** (Spectrally small within-row nonuniformity)**.** *Consider the setting above (Equation (3.3)). Suppose the assumptions of Theorem 3.2 hold, $K, r, \mu = O(1)$, $q = \Omega(1)$, and $m \geq n$. Then, with probability at least $1 - O(1/(m+n))$, it holds that $(1/\sqrt{m}) \cdot \|\widehat{M} - M\|_{2,\infty} \leq \nu\, \widetilde{O}(n^{-1/2})$.*

Thus, Corollary 3.5 preserves the $\widetilde{O}(n^{-1/2})$ row-normalized rate of the row-homogeneous case, with $\nu$ inflating only the constant. Crucially, the algorithm itself does not change between the two regimes; the analyst does not need to verify the spectral bound or estimate $\nu$.

## 4 Technical Overview

In this section, we sketch the key ideas behind the proof of Theorem 3.2. In this section, we focus on $a = 1$ and, hence, omit it from the notation, writing, *e.g.*, $A$ and $P$ for $A(1)$ and $P(1)$ respectively. Now, our goal is to prove that, with probability at least $1 - O(1/(m+n))$, $\|\widehat{A} - A\|_{2,\infty}$ is bounded, up to a universal constant, by the right-hand side of Equation (3.1). Here, the constants $K$, $r$, and $\mu$ are from Assumption 3.1 and the parameters $(r_q, q, P)$ are as defined in Equation (1.2). Before presenting the proof, we need to set up some notation.

**Notation and basic observations.** We use $A^{ub}$ to denote $\mathbb{E}\,\widetilde{Y}^{ub}$, where $\widetilde{Y}^{ub}$ is from Algorithm 1. A direct computation shows $A^{ub} := (p_{ij} a_{ij}/p_i)_{ij}$. For any $X \in \mathbb{R}^{n\times m}$, we denote the corresponding "observed" matrix by $\widetilde{X} = (\widetilde{x}_{ij})_{ij}$ and the corresponding "scaled" matrix by $\widetilde{X}^{ub} := (\widetilde{x}_{ij}/p_i)_{ij}$. With this notation, since $Y = A + E$, we can write $\widetilde{Y}^{ub} = \widetilde{A}^{ub} + \widetilde{E}^{ub}$. Finally, we define

$$E_0 := A^{ub} - A = \left(\left(\frac{p_{ij}}{p_i} - 1\right) a_{ij}\right)_{ij} \qquad \text{and} \qquad E_R := (\widetilde{Y}^{ub} - A^{ub}) + \widetilde{E}^{ub}\,.$$

Next, we observe that Assumption 3.1(5) and the choice that $\widetilde{\sigma}^{\mathrm{ub}}_s - \widetilde{\sigma}^{\mathrm{ub}}_{s+1} \gtrsim KT(a)$ are equivalent to $\sigma_1 > 10r(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}})$, and $\widetilde{\sigma}^{\mathrm{ub}}_s - \widetilde{\sigma}^{\mathrm{ub}}_{s+1} \geq 8(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}})$ respectively. (Where we selected the constants to simplify exposition.)

**Our approach.** At a high level, to prove (3.1), we use the triangle inequality $\|\widehat{A} - A\|_{2,\infty} \leq \|A_s - A\|_{2,\infty} + \|\widehat{A} - A_s\|_{2,\infty}$ and bound the two terms separately. Term 1, $\|A_s - A\|_{2,\infty}$, is the tail error of the low-rank approximation; we bound it using the singular value decomposition of $A$ and the definition of the threshold position $s$. Term 2, $\|\widehat{A} - A_s\|_{2,\infty}$, measures the perturbation of low-rank approximations in $\|\cdot\|_{2,\infty}$; bounding it is the key difficulty.

**Step 1 (Bounding Term 1).** Since $A = \sum_{i=1}^r \sigma_i u_i v_i^\top$, $A - A_s = \sum_{i=s+1}^r \sigma_i u_i v_i^\top$. Therefore,

$$\|A_s - A\|_{2,\infty} = \max_{1\leq k\leq n} \|e_{n,k}^\top (A_s - A)\|_2 = \max_{1\leq k\leq n} \|\sum\nolimits_{r\geq i>s} \sigma_i u_{ik} v_i^\top\|_2\,.$$

Since the right singular vectors are orthonormal and $\sigma_i \leq \sigma_{s+1}$ for $i > s$, we obtain

$$\|A_s - A\|_{2,\infty} = \max_{1\leq k\leq n} \sqrt{\sum\nolimits_{i=s+1}^r \sigma_i^2 u_{ik}^2} \leq \max_{s<i\leq r} \|u_i\|_\infty \sqrt{\sum\nolimits_{i=s+1}^r \sigma_i^2} \leq \max_{s<i\leq r} \|u_i\|_\infty \sqrt{r}\sigma_{s+1}\,.$$

We claim that $\sigma_{s+1} \leq 10r(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}})$. Suppose, for contradiction, that $\sigma_{s+1} >$

$10r(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}})$. Since $\mathrm{rank}(A) \le r$ (*i.e.*, $\sigma_{r+1} = 0$), it follows that $s+1 \le r$. Arguing as in Remark B.2 (with $\sigma_{s+1}$ in place of $\sigma_1$), there exists some $j$ with $s+1 \le j \le r$ such that $\widetilde{\sigma}^{ub}_j - \widetilde{\sigma}^{ub}_{j+1} > 8(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}})$. This contradicts the definition of $s$ as the largest index satisfying $\widetilde{\sigma}^{ub}_s - \widetilde{\sigma}^{ub}_{s+1} > 8(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}})$. Hence, $\sigma_{s+1} \le 10r(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}})$. Therefore,

$$\|A_s - A\|_{2,\infty} \le 10r^{3/2}(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}}) \cdot \max_{s<i\le r} \|u_i\|_\infty .$$

To bound $\|E_R\|$, we use the following lemma. Its proof will be presented later in Section D.3.

**Lemma 4.1.** *With high probability,* $\|E_R\|_{\mathrm{op}} \le 12K\sqrt{\log(m+n)} \cdot \sqrt{r_p(m+n)/q}$.

We also have $\|E_0\|_{\mathrm{op}} \le K_A\|P\|_{\mathrm{op}}$. By the definition of $\mu$, $\max_{s<i\le r} \|u_i\|_\infty \le \mu(1/\sqrt{m} + 1/\sqrt{n})$. Combining all estimates, we obtain

$$\|A_s - A\|_{2,\infty} \le 12\sqrt{\log(m+n)}Kr^{3/2}\mu\left(\sqrt{m}+\sqrt{n}\right) \cdot \left(\sqrt{\frac{r_p}{mq} + \frac{r_p}{nq}} + \frac{\|P\|_{\mathrm{op}}}{\sqrt{mn}}\right) . \tag{4.1}$$

**Step 2 (Bounding Term 2).** Our estimate uses the following theorem, a key technical contribution. We compare its proof techniques with prior work in Section 4.1. Further discussion of perturbations of low-rank approximations in $\|\cdot\|_{2,\infty}$ and the full proof appear in Section C.

**Theorem 4.2.** *Let* $\widetilde{A} = A + E_R + E_0$, *where* $E_R$ *is* $(K,\sigma)$*-bounded. There is a universal constant* $C > 0$ *satisfying: If* $\delta_s \ge 6(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}})$, *then with probability* $1 - O(1/(m+n))$,

$$\frac{\|\widetilde{A}_s - A_s\|_{2,\infty}}{C\sqrt{r}} \le \left(\log^2(m+n) + \frac{K\log^4(m+n)}{\sqrt{m+n}}\right)\left(\frac{\mu}{\sqrt{m}} + \frac{\mu}{\sqrt{n}}\right)\frac{\sigma_s\sqrt{m+n}\,(\sigma + \|E_0\|_{\mathrm{op}})}{\delta_s} .$$

Since $\widetilde{Y}^{ub} = A + E_R + E_0$, we have the following observations.

- Since $\widetilde{\sigma}^{ub}_s - \widetilde{\sigma}^{ub}_{s+1} \ge 8(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}})$, by Weyl's inequality (Theorem B.3), we have

$$\sigma_s - \sigma_{s+1} \;\ge\; \widetilde{\sigma}^{ub}_s - \widetilde{\sigma}^{ub}_{s+1} - 2(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}}) \;\ge\; 6(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}}) .$$

- By the definition of $E_R = (E_{Rij})_{ij}$, we have

$$E_{Rij} = \begin{cases} (1-p_{ij})a_{ij}/p_i + \varepsilon_{ij}/p_i \text{ with probability } p_{ij} , \\ -p_{ij}a_{ij}/p_i \text{ with probability } 1-p_{ij}. \end{cases}$$

For each $\ell \ge 2$, a direct moment calculation gives $\mathbb{E}\left[|E_{Rij}|^\ell\right] \le (p_{ij}/p_i^\ell) \cdot (K_A + K_E)^\ell$. Since $K = K_A + K_E$, the right-hand side equals $(p_{ij}/p_i^\ell) \cdot K^\ell$. These observations allow us to apply Theorem 4.2 on the pair $(\widetilde{Y}^{ub}, A)$ with the noises $E_R, E_0$. Indeed, the parameters $K$ and $\sigma$ in Theorem 4.2 are replaced respectively by $K/q$ and $\sqrt{r_pK/q}$ in our setting. Further, $\sigma_s/\delta_s = 1 + (\sigma_{s+1}/\delta_s) \le 1 + \frac{r(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}})}{8(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}})} \le 2r$. Thus, since $\widehat{A} = \widetilde{Y}^{ub}_s$, with probability at least $1 - O((m+n)^{-1})$,

$$\begin{aligned} \|\widehat{A} - A_s\|_{2,\infty} &\lesssim \sqrt{r}\,\frac{\sigma_s}{\delta_s}\,\log^4(m+n)K\mu(\sqrt{m}+\sqrt{n}) \cdot \left[\sqrt{\frac{r_p}{mq} + \frac{r_p}{nq}} + \frac{\|P\|_{\mathrm{op}}}{\sqrt{mn}}\right] \\ &\lesssim \log^4(m+n)Kr^{3/2}\mu(\sqrt{m}+\sqrt{n}) \cdot \left[\sqrt{\frac{r_p}{mq} + \frac{r_p}{nq}} + \frac{\|P\|_{\mathrm{op}}}{\sqrt{mn}}\right] . \end{aligned} \tag{4.2}$$

Combining (4.1) and (4.2) yields Theorem 3.2.

### 4.1 High-level Proof Sketch of Theorem 4.2

To prove Theorem 4.2, we adapt the contour integral approach used in recent works; *e.g.*, Tran & Vu (2025a; 2026a; 2025b); Koltchinskii & Xia (2016); O'Rourke et al. (2023). We analyze $\widetilde{A}_s - A_s$ through the symmetrized matrices:

$$\mathcal{A} := \begin{bmatrix} 0 & A \\ A^\top & 0 \end{bmatrix}, \qquad \mathcal{E} := \begin{bmatrix} 0 & E \\ E^\top & 0 \end{bmatrix}, \qquad \widetilde{\mathcal{A}} := \begin{bmatrix} 0 & \widetilde{A} \\ \widetilde{A}^\top & 0 \end{bmatrix} .$$

The symmetrized versions are related to $\widetilde{A}_s - A_s$ by

$$\|\widetilde{A}_s - A_s\|_{2,\infty} := \max_{1\le k\le n} \|e_{n,k}^\top(\widetilde{A}_s - A_s)\|_2 = \max_{1\le k\le n} \|e_{m+n,k}^\top(\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s})\|_2 .$$

Thus, to prove Theorem 4.2 it suffices to bound $\|e_1^\top(\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s})\|_2$. Next, using the Cauchy integral theorem (Theorem B.4), we obtain

$$e_1^\top(\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s}) = \tfrac{1}{2\pi \mathbf{i}} \textstyle\int_\Gamma z \cdot e_1^\top[(zI - \widetilde{\mathcal{A}})^{-1})^{-1} - (zI - \mathcal{A})^{-1}]\,\mathrm{d}z ,$$

where $\Gamma$ is a contour in $\mathbb{C}$ that encloses $\pm\sigma_1, \pm\sigma_2, \ldots, \pm\sigma_s$ and excludes $\pm\sigma_{s+1}, \pm\sigma_{s+2}, \ldots, \pm\sigma_r$.

**Challenge (Prior techniques are insufficient to bound $\|\cdot\|_{2,\infty}$ norm).** If instead of the $\|\cdot\|_{2,\infty}$, we wanted to bound the $\|\cdot\|_2$ or $\|\cdot\|_{\mathrm{op}}$ norms, then one could straightforwardly bound $\|e_1^\top(\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s})\|_2$ by $\|\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s}\|_{\mathrm{op}}$, and then apply the existing spectral-norm bounds on perturbations of low-rank approximations (*e.g.*, Tran et al. (2025); Tran & Vu (2025c); Eckart & Young (1936)). However, to obtain meaningful bounds on the Heterogenous treatment effect, we need to focus on the $\|\cdot\|_{2,\infty}$ norm and, here, using the aforementioned bounds yields a suboptimal bound which can off by a large factor, of up to $\sqrt{n+m}$. Thus, obtaining a sharp bound with respect to the row-wise $\ell_2$-norm, which is crucial to prove Theorem 4.2, remains a formidable analytical challenge that requires some new ideas.

**Ideas.** For the sharp bound, we adapt the contour expansion method of Tran & Vu (2026a; 2025a); Tran et al. (2025). Repeatedly applying the Sherman–Morrison–Woodbury formula $M^{-1} - (M+N)^{-1} = (M+N)^{-1}NM^{-1}$ (Horn & Johnson, 2012) with $\widetilde{A} = A + E$ gives

$$e_1^\top(\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s}) = \textstyle\sum_{k=1}^\infty e_1^\top H_k, \quad \text{where} \quad H_k := \tfrac{1}{2\pi\mathbf{i}} \textstyle\int_\Gamma z \cdot (zI - \mathcal{A})^{-1}[\mathcal{E}(zI - \mathcal{A})^{-1}]^k\,\mathrm{d}z .$$

Intuitively, Tran et al. (2025) used a contour bootstrapping argument to show that $\|\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s}\|_{\mathrm{op}}$ is of the same order as $\|H_1\|_{\mathrm{op}}$. For entrywise control, Tran & Vu (2025a) bounded $\|H_k\|_\infty$ for all $k \ge 1$. In our row-wise $\ell_2$ setting, we instead bound $\|e_1^\top H_k\|_2$ for $1 \le k \le \ell \log n$, with a suitable constant $\ell$, and show that the remaining tail is negligible.

**Further challenges.** Bounding $\|e_1^\top H_k\|_2$ presents several further challenges. Sharp estimates require controlling the interactions $\left|e_1^\top \mathcal{E}^\ell \binom{U}{V}\right|, \forall 1 \le \ell \le k$. Naive bounds based on $\|e_1^\top \mathcal{E}^\ell\|_2$, $\|\mathcal{E}\|_{\mathrm{op}}^\ell$, or $\left\|\mathcal{E}^\ell \binom{U}{V}\right\|_2$ are suboptimal. Since $\mathcal{E}$ is random, one expects $\mathcal{E}^\ell$ to spread mass across the entries of $\binom{U}{V}^\top$, making the incoherence parameter $\mu$ essential. In Tran & Vu (2025a), the authors handled $\left|e_1^\top \mathcal{E}^\ell \binom{U}{V}\right|$ when $\mathcal{E}$ is random and mean-zero. In our setting, $\mathcal{E} = \mathcal{E}_R + \mathcal{E}_0$ contains both random and deterministic components, so separating their contributions is highly nontrivial; see Section D.

## 5 Experimental Results

Next, we describe the experimental evaluation of our estimator; Section A gives the details. We run two types of experiments. First, on synthetic panels, we test whether the worst-row error decays at the rate predicted by Corollary 3.3. Second, on synthetic panels and on semi-synthetic panels calibrated to the HeartSteps mobile-health trial (Klasnja et al., 2019), we estimate each unit's average effect over windows of $k$ consecutive times and compare with direct Horvitz–Thompson (HT) and Hájek estimates. In both, the true effects are known. For a fixed aspect ratio, the median normalized worst-row error $\|\widehat{M} - M\|_{2,\infty}/\sqrt{m}$ scales as $n^{-0.503}$, close to the predicted $n^{-1/2}$. For window averages, our estimator is most accurate on short windows, where direct estimates use few observations. At $k = 1$ on synthetic panels, its median RMSE is $0.18$ versus $0.60$ for HT (Hájek is undefined). At $k = 8$ on HeartSteps-calibrated panels, it is $0.83$ versus $3.01$ for HT and $2.27$ for Hájek. The gap narrows as windows grow, and Hájek is slightly more accurate for the longest windows. The semi-synthetic effects are constant within each participant, so these panels test averaging over windows, not recovery of time-varying effects.

## 6 Conclusion

We give a row-scaled truncated-SVD estimator for unit-specific treatment effects under unknown, nonuniform propensities, using only the observed outcomes and assignments. A new row-wise perturbation bound yields guarantees for averages over sufficiently large subsets of times; in experiments, the gains over direct estimators are largest for short windows. Future directions include weakening the low-rank, incoherence, and signal-to-noise assumptions, extending the analysis to adaptive or dependent assignments, and sharpening the dependence on design nonuniformity.

## AI USE STATEMENT

Generative AI was used to assist in writing code for the experiments. Generative AI tools also assisted with editing the exposition and formatting the manuscript. The authors are responsible for the paper's content, including material developed with AI assistance.

## REFERENCES


Alberto Abadie and Javier Gardeazabal. The economic costs of conflict: A case study of the Basque Country. *American Economic Review*, 93(1):113–132, 2003. doi: 10.1257/000282803321455188.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505, 2010. doi: 10.1198/jasa.2009.ap08746.

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics*, 48(3):1452, 2020.

Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. On robustness of principal component regression. *Journal of the American Statistical Association*, 116(536):1731–1745, 2021. doi: 10.1080/01621459.2021.1928513.

Anish Agarwal, Munther Dahleh, Devavrat Shah, and Dennis Shen. Causal matrix completion. In *Proceedings of the Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 3821–3826. PMLR, 2023. URL https://proceedings.mlr.press/v195/agarwal23c.html.

Anish Agarwal, Devavrat Shah, and Dennis Shen. On model identification and out-of-sample prediction of PCR with applications to synthetic controls. *Journal of Machine Learning Research*, 26(117):1–58, 2025a. URL https://jmlr.org/papers/v26/23-0102.html.

Anish Agarwal, Devavrat Shah, and Dennis Shen. Synthetic interventions: Extending synthetic controls to multiple treatments. *Operations Research*, 74(2):840–859, 2025b. doi: 10.1287/opre.2025.1590.

Muhammad Amjad, Devavrat Shah, and Dennis Shen. Robust synthetic control. *Journal of Machine Learning Research*, 19(22):1–51, 2018. URL https://jmlr.org/papers/v19/17-777.html.

Muhammad Amjad, Vishal Misra, Devavrat Shah, and Dennis Shen. mRSC: Multi-dimensional robust synthetic control. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3(2):1–27, 2019. doi: 10.1145/3341617.3326152.

Dmitry Arkhangelsky, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager. Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118, 2021. doi: 10.1257/aer.20190159.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. doi: 10.1073/pnas.1510489113.

Susan Athey, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021. doi: 10.1080/01621459.2021.1891924.

Jushan Bai and Serena Ng. Matrix completion, counterfactuals, and factor analysis of missing data. *Journal of the American Statistical Association*, 116(536):1746–1763, 2021. doi: 10.1080/01621459.2021.1967163.

Afonso S Bandeira and Ramon van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Ann. Probab.*, 44(4):2479–2506, 2016.

Abhinav Bhardwaj and Van Vu. Matrix perturbation: Davis-Kahan in the infinity norm. *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 880–934, 2024.

Iavor Bojinov, David Simchi-Levi, and Jinglong Zhao. Design and analysis of switchback experiments. *Management Science*, 69(7):3759–3777, 2023. doi: 10.1287/mnsc.2022.4583.

Audrey Boruvka, Daniel Almirall, Katie Witkiewitz, and Susan A. Murphy. Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 113 (523):1112–1121, 2018. doi: 10.1080/01621459.2017.1305274.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009. doi: 10.1007/s10208-009-9045-5.

Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010. doi: 10.1109/TIT.2010.2044061.

Joshua Cape, Minh Tang, and Carey E Priebe. The two-to-infinity norm and singular subspace geometry with applications to high-dimensional statistics. *Annals of Statistics*, 47(5):2405–2439, 2019.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015. doi: 10.1214/14-AOS1272.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

C. Dwork, K. Talwar, A. Thakurta, and L. Zhang. Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, pp. 11–20, 2014.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Iván Fernández-Val, Hugo Freeman, and Martin Weidner. Low-rank approximations of nonseparable panel models. *The Econometrics Journal*, 24(2):C40–C77, 2021. doi: 10.1093/ectj/utab007.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4 edition, 2013. ISBN 9781421407944.

Miguel A. Hernan and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC monographs on statistics & applied probability. Taylor & Francis, 2023. ISBN 9781315374932. URL https://books.google.co.in/books?id=FPkN0AEACAAJ.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, 2008. See §1.3 and §7.6 for the Cauchy–Dunford integral.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Always valid inference: Continuous monitoring of A/B tests. *Operations Research*, 70(3):1806–1821, 2022. doi: 10.1287/opre.2021.2135.

Tosio Kato. *Perturbation Theory for Linear Operators*. Classics in Mathematics. Springer, New York, NY, 1980.

Predrag Klasnja, Eric B. Hekler, Saul Shiffman, Audrey Boruvka, Daniel Almirall, Ambuj Tewari, and Susan A. Murphy. Microrandomized trials: An experimental design for developing just-in-time adaptive interventions. *Health Psychology*, 34(S):1220–1228, 2015. doi: 10.1037/hea0000305.

Predrag Klasnja, Shawna Smith, Nicholas J. Seewald, Andy Lee, Kelly Hall, Brook Luers, Eric B. Hekler, and Susan A. Murphy. Efficacy of contextually tailored suggestions for physical activity: A micro-randomized optimization trial of HeartSteps. *Annals of Behavioral Medicine*, 53(6):573–582, 2019. doi: 10.1093/abm/kay067.

V. Koltchinskii and D. Xia. Perturbation of linear forms of singular vectors under Gaussian noise. In *High Dimensional Probability VII: The Cargèse Volume*, volume 71 of *Progress in Probability*, pp. 397–423. Birkhäuser, Cham, 2016. doi: 10.1007/978-3-319-40519-3_18.

Richard L. Kravitz, Naihua Duan, and Joel Braslow. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly*, 82(4):661–687, 2004. doi: 10.1111/j.0887-378X.2004.00327.x.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 2019. doi: 10.1073/pnas.1804597116.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 661–670. ACM, 2010. doi: 10.1145/1772690.1772758.

Peng Liao, Predrag Klasnja, Ambuj Tewari, and Susan A. Murphy. Sample size calculations for micro-randomized trials in mHealth. *Statistics in Medicine*, 35(12):1944–1971, 2016. doi: 10.1002/sim.6847.

Oren Mangoubi and Nisheeth K. Vishnoi. Private low-rank approximation for covariance matrices, Dyson Brownian Motion, and eigenvalue-gap bounds for Gaussian perturbations. *J. ACM*, 72(2), March 2025. ISSN 0004-5411. doi: 10.1145/3716496. URL https://doi.org/10.1145/3716496.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(80):2287–2322, 2010.

Sean O'Rourke, Van Vu, and Ke Wang. Matrices with Gaussian noise: Optimal estimates for singular subspace perturbation. *IEEE Transactions on Information Theory*, 2023.

Tianchen Qian, Ashley E. Walton, Linda M. Collins, Predrag Klasnja, Stephanie T. Lanza, Inbal Nahum-Shani, Mashfiqui Rabbi, Michael A. Russell, Maureen A. Walton, Hyesun Yoo, and Susan A. Murphy. The micro-randomized trial for developing digital interventions: Experimental design and data analysis considerations. *Psychological Methods*, 27(5):874–894, 2022. doi: 10.1037/met0000283.

Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1670–1679. PMLR, 2016.

G.W. Stewart and J. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

Adith Swaminathan and Thorsten Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, 16(52):1731–1755, 2015.

BaoLinh Tran and Van Vu. Fast exact recovery of noisy matrix from few entries: the infinity norm approach. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025a.

Phuc Tran and Nisheeth K. Vishnoi. Perturbation bounds for low-rank inverse approximations under noise. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025.

Phuc Tran and Van Vu. Davis–Kahan theorem under a moderate gap condition. *Communications in Contemporary Mathematics*, 2025b. World Scientific, doi: 10.1142/S021919972550035X.

Phuc Tran and Van Vu. Matrices perturbed by random noise: The accuracy of low-rank approximation. 2025c. arXiv:2511.08875.

Phuc Tran and Van Vu. New matrix perturbation bounds with relative strength: Perturbation of eigenspaces. *arXiv preprint: 2409.20207*, 2026a.

Phuc Tran and Van Vu. Regional stability and new eigenvalue perturbation bounds, 2026b. URL https://arxiv.org/abs/2609.14303. arXiv:2609.14303.

Phuc Tran, Nisheeth K. Vishnoi, and Van H. Vu. Spectral perturbation bounds for low-rank approximation with applications to privacy. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015. doi: 10.1561/2200000048.

Van Vu. Spectral norm of random matrices. *Combinatorica*, 27(6):721–736, 2007.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

Hermann Weyl. Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen. *Mathematische Annalen*, 71(4):441–479, 1912.

Yiqing Xu. Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis*, 25(1):57–76, 2017. doi: 10.1017/pan.2016.2.

# A Empirical Evaluation

We evaluate our estimator on synthetic panels and on semi-synthetic panels calibrated to the HeartSteps study. In both settings, the potential-outcome means are known, so we can measure estimation error directly. We run two types of experiments. First, on synthetic panels, we measure how the worst-row error changes with the panel size and compare it with the rate in Corollary 3.3. Second, in both settings, we estimate unit-specific treatment effects averaged over time windows of different lengths and compare with direct estimators, which use fewer observations as the windows become shorter. Section A.1 describes the estimators and metrics, and Sections A.2 and A.3 present the results. In summary, the synthetic error decays at close to the predicted rate, and in both settings our estimator is most accurate, relative to direct estimators, on short windows.

## A.1 Estimators and Evaluation Metrics

**Our estimator.** We use a practical variant of Algorithm 1 that selects the truncation rank by cross-validation instead of using the theoretical singular-gap threshold. For each action $a \in \{0, 1\}$, we zero-fill the unobserved entries, divide each row by its empirical observation frequency (lower-bounded by $1/m$), and compute a truncated SVD. We choose the rank separately for the two actions from $\{0, \dots, 10\}$, and subtract the resulting estimates to obtain $\widehat{M}$.

**Rank selection.** We select each rank by five-fold cross-validation on the observed entries, using the one-standard-error rule. For each action, we shuffle each row's observed entries and assign them round-robin to the five folds; all candidate ranks use the same folds. If a row has fewer than two observations for an action, we keep them in the training data in every fold. For each fold, we fit on the remaining observations and compute the mean squared error on the held-out factual outcomes. In this fit, the row frequency is the number of training observations for that action divided by $m$, lower-bounded by $1/m$, so a row with no training observations stays zero. We average the losses over the nonempty folds and select the smallest rank whose average loss is within one standard error of the minimum. Here, the standard error is the sample standard deviation of the fold losses at the minimizing rank, divided by the square root of the number of nonempty folds. We then refit on all factual observations; rank zero gives the zero matrix. Rank selection uses neither a separate validation panel nor the true means.

**Direct estimators.** For a unit $i$ and a window $S \subseteq [m]$, the target is $\mathrm{Avg}_i(S) = {}^1/_{|S|} \sum_{j \in S} M_{ij}$. Our estimate averages the corresponding entries of $\widehat{M}$. We compare it with direct Horvitz–Thompson (HT) and Hájek estimates. In all experiments below, treatment is independently assigned with a common probability $p$. The HT estimate uses this known probability:

$$\widehat{\mathrm{Avg}}_i^{\mathrm{HT}}(S) := \frac{1}{|S|} \sum_{j \in S} \left( \frac{D_{ij} Y_{ij}^{\mathrm{obs}}}{p} - \frac{(1 - D_{ij}) Y_{ij}^{\mathrm{obs}}}{1 - p} \right).$$

The Hájek estimate is the difference between the treated and control sample means within the window. It is defined only when the window contains observations from both actions. For full-matrix recovery, we also compare with the untruncated row-scaled estimate, obtained by omitting the SVD truncation, and with the cellwise HT estimate.

**Metrics.** For full-matrix recovery, we report normalized worst-row error $\|\widehat{M} - M\|_{2,\infty} / \sqrt{m}$. For temporal queries, we partition each row into nonoverlapping contiguous windows of length $k$, fit the estimator once per panel, and reuse it for every window length. All reported lengths divide $m$. Within each repetition, we compute the root mean squared error (RMSE) over all $n(m/k)$ unit-window queries. We then report the median across repetitions and, for our estimator, the empirical 10th and 90th percentiles. These percentiles describe variation across simulated panels, rather than confidence intervals for the median.

Hájek RMSE is computed only over windows on which the estimate is defined; our estimator and HT are evaluated over all windows. We report Hájek coverage, the fraction of unit-window queries on which both actions occur, to make this difference explicit. Comparisons within a repetition use the same outcomes and assignments.

### A.2 Synthetic Panels

**Data generation.** For each action $a \in \{0, 1\}$, we generate a rank-3 mean matrix $A(a) = U(a)\,\mathrm{diag}(c(a))V^\top$. The two matrices $U(a) \in \{-1, 1\}^{n\times 3}$ are drawn independently, with independent Rademacher entries, and redrawn if they do not have full column rank. The shared temporal factors $V \in \mathbb{R}^{m\times 3}$ have a constant component, a change halfway through the panel, and an alternating component. With time indexed from 1, they are

$$V_{j1} = 1, \qquad V_{j2} = \begin{cases} 1, & j \le \lceil m/2 \rceil, \\ -1, & j > \lceil m/2 \rceil, \end{cases} \qquad V_{j3} = (-1)^{j-1}.$$

We set $c(0) = (0.28, 0.20, 0.12)$ and $c(1) = (0.32, 0.16, 0.10)$. Thus the signal entries remain bounded as the dimensions grow. We independently draw noise entries $E_{ij}(a)$ uniformly from $[-0.25, 0.25]$ and assignments $D_{ij} \sim \mathrm{Ber}(1/2)$. The observed outcome is $Y_{ij}^{\mathrm{obs}} = D_{ij}(A_{ij}(1) + E_{ij}(1)) + (1 - D_{ij})(A_{ij}(0) + E_{ij}(0))$, and the target is $M = A(1) - A(0)$.

**Dependence on panel size.** We generate 200 independent panels at each of the seven dimension pairs in Table 1, keeping $m/n \approx 210/44$, the aspect ratio of the HeartSteps study. The table compares normalized worst-row error with the dimension-dependent term $\sqrt{2(1/n + n/m^2)}$ from Corollary 3.3: here $q = 1/2$, $r_p = 1$, and $P(0) = P(1) = 0$. This term is shown without fitted rescaling. It omits the theorem's multiplicative constants, structural factors, and logarithmic factors, so it serves as a guide to the rate rather than a numerical error bound.

Table 1: Synthetic dimension scaling. Normalized worst-row error over 200 panels per size; brackets contain the 10th and 90th percentiles. The final column is the dimension-dependent rate term, not the full theorem bound.

| $n$ | $m$ | Our estimator | $\sqrt{2(1/n + n/m^2)}$ |
|---|---|---|---|
| 16 | 76 | 0.388 [0.348, 0.440] | 0.361 |
| 24 | 115 | 0.339 [0.301, 0.385] | 0.295 |
| 32 | 153 | 0.289 [0.262, 0.335] | 0.255 |
| 44 | 210 | 0.249 [0.227, 0.279] | 0.218 |
| 64 | 305 | 0.199 [0.183, 0.218] | 0.181 |
| 96 | 458 | 0.163 [0.151, 0.178] | 0.147 |
| 128 | 611 | 0.141 [0.131, 0.155] | 0.128 |

The median error decreases from 0.388 at $(n, m) = (16, 76)$ to 0.141 at $(128, 611)$. A least-squares fit of log median error against $\log n$ gives slope $-0.503$, close to the $-1/2$ dependence predicted when the aspect ratio is fixed. Since $n$ and $m$ grow together, this experiment does not separately identify their effects.

**Unit-specific effects at different temporal resolutions.** We next fix $(n, m) = (44, 210)$, generate a separate set of 200 panels, and evaluate the window lengths in Table 2. Our estimator has lower median RMSE than HT at every length, and the gap is largest for short windows: at $k = 1$, the errors are 0.178 and 0.601, respectively. Our estimator also has lower median RMSE than Hájek for $2 \le k \le 42$, although Hájek is evaluated only on windows where it is defined (Section A.1). For the longest windows, $k = 70$ and $k = 210$, Hájek is slightly more accurate. Thus, low-rank structure helps most for localized queries, where each direct estimate uses few observations, and its advantage shrinks for broad averages.

Table 2: Synthetic window queries, $(n, m) = (44, 210)$. Entries are median RMSEs over 200 panels; brackets contain the 10th and 90th percentiles for our estimator. Hájek uses only defined windows, and coverage is pooled across repetitions.

| Window | Our estimator | HT | Hájek | Coverage |
|---|---|---|---|---|
| 1 | 0.178 [0.171, 0.187] | 0.601 | – | 0.0% |
| 2 | 0.124 [0.116, 0.130] | 0.424 | 0.257 | 50.0% |
| 5 | 0.088 [0.083, 0.094] | 0.268 | 0.169 | 93.8% |
| 10 | 0.070 [0.065, 0.077] | 0.190 | 0.115 | 99.8% |
| 21 | 0.061 [0.055, 0.068] | 0.130 | 0.074 | 100.0% |
| 42 | 0.052 [0.047, 0.059] | 0.092 | 0.055 | 100.0% |
| 70 | 0.048 [0.042, 0.055] | 0.071 | 0.044 | 100.0% |
| 210 | 0.031 [0.027, 0.036] | 0.041 | 0.029 | 100.0% |

### A.3 HeartSteps-Calibrated Panels

**Data and calibration.** The HeartSteps micro-randomized trial studied contextually tailored activity suggestions (Klasnja et al., 2019). Since the alternative potential outcome is unobserved at each decision point, the study data do not directly reveal treatment-effect estimation error. We therefore construct panels with known potential-outcome means, using HeartSteps to set the outcome scale, participant variation, and noise distribution.

**Baseline outcomes.** We use the public HeartSteps V1 data (`suggestions.csv`)[2] for 37 participants. Within each participant's first 210 planned decision points, we retain the first 104 available opportunities, giving a $37 \times 104$ panel. Column $j$ records the $j$th available opportunity for each participant, rather than a common calendar time. We take the pre-decision step counts (the dataset's `jbsteps30pre.zero` field), transform them by $\log(\text{steps} + 0.5)$, and winsorize entries above the pooled 99th percentile. Let $B$ denote this transformed panel. We set $A(0) = B_3$, its rank-3 truncated SVD, without centering $B$.

**Treatment effects.** We construct a participant-specific effect by ranking participants by their row-average activity in $B$. If $r_i \in \{0, \ldots, 36\}$ is participant $i$'s rank in increasing order, set $h_i = 0.75 + 0.5 r_i/36$ and $M_{ij} = h_i \log(1.14)$. Thus, the mean effect is $\log(1.14) \approx 0.131$, calibrated to the reported 14% increase in step count from an activity suggestion, averaged across study days and suggestion types (Klasnja et al., 2019). Effects range from approximately $0.098$ to $0.164$ and are constant across each participant's opportunities. The heterogeneity is imposed by the simulation, rather than estimated from the trial. We set $A(1) = A(0) + M$, which has rank at most 4.

**Noise and assignments.** To generate noise, we pool the residuals $B - B_3$, subtract their mean, clip them symmetrically at the 99th percentile of their absolute values, and center them again. Each noise entry $E_{ij}(a)$ is drawn independently with replacement from this pool, and we set $Y_{ij}(a) = A_{ij}(a) + E_{ij}(a)$ without further scaling, rounding, or clipping. We generate fresh assignments independently with probability $p = 0.6$, and reveal only the factual outcomes and assignments to the estimators. We repeat this experiment 500 times, keeping the calibrated means and residual pool fixed while resampling the noise, assignments, and cross-validation folds. This construction uses the calibrated residual distribution but does not retain dependence across participants or opportunities in the noise.

**Unit-specific effects.** Table 3 reports window-query errors using the same estimators and metrics as in Section A.1. Our estimator has lower median RMSE than HT at every reported window length, and lower median RMSE than Hájek for lengths 2 through 52. For example, at length 8, the respective errors are 0.834, 3.009, and 2.274. At length 104, Hájek is slightly more accurate: its median RMSE is 0.581, compared with 0.648 for our estimator. Hájek is again evaluated only on defined windows, with coverage reported in the table.

[2] https://github.com/klasnja/HeartStepsV1.

Table 3: HeartSteps-calibrated window queries, $(n, m) = (37, 104)$. Entries are median RMSEs over 500 repetitions; brackets contain the 10th and 90th percentiles for our estimator. Hájek uses only defined windows. Coverage is pooled across repetitions and rounded to one decimal place.

| Window | Our estimator | HT | Hájek | Coverage |
|---|---|---|---|---|
| 1 | 1.699 [1.579, 1.850] | 8.520 | – | 0.0% |
| 2 | 1.285 [1.156, 1.424] | 6.026 | 4.093 | 48.0% |
| 4 | 1.005 [0.893, 1.139] | 4.258 | 3.189 | 84.5% |
| 8 | 0.834 [0.730, 0.963] | 3.009 | 2.274 | 98.2% |
| 13 | 0.769 [0.664, 0.894] | 2.368 | 1.737 | 99.9% |
| 26 | 0.707 [0.611, 0.816] | 1.674 | 1.192 | 100.0% |
| 52 | 0.670 [0.580, 0.779] | 1.181 | 0.827 | 100.0% |
| 104 | 0.648 [0.562, 0.765] | 0.834 | 0.581 | 100.0% |

**Full-matrix recovery.** Table 4 compares normalized worst-row errors. The median error is 2.835 for our estimator, compared with 11.699 for the untruncated row-scaled estimate and 11.618 for cellwise HT. Thus, low-rank regularization reduces this error by approximately a factor of four relative to these two unsmoothed estimates.

Table 4: HeartSteps-calibrated full-matrix recovery. Normalized worst-row error over 500 repetitions; brackets contain the 10th and 90th percentiles.

| Method | Median [10th, 90th percentile] |
|---|---|
| Our estimator | 2.835 [2.511, 3.414] |
| Untruncated row-scaled | 11.699 [11.176, 12.363] |
| Cellwise HT | 11.618 [11.097, 12.231] |

**Limitations.** These comparisons show lower error than the direct estimators, particularly for short windows. However, the remaining errors are large relative to the injected effects: our estimator's smallest median RMSE in Table 3, 0.648, is about five times the mean effect $\log(1.14) \approx 0.131$. Moreover, each participant's true effect is constant over opportunities, so this experiment tests averaging at different window lengths, not recovery of time-varying effects. Finally, since the counterfactual outcomes are simulated, this experiment does not validate counterfactual accuracy on the original HeartSteps outcomes.

**Reproducibility.** The synthetic scaling and window experiments use seed values 910260725 and 950260725, respectively; each uses 200 repetitions per setting. The HeartSteps experiment uses seed 20260726 and 500 repetitions. All reported results are computed from the stored repetition-level errors; the truncation ranks are selected without access to the true treatment-effect matrices.

## B Additional Discussion and Preliminaries

In this section, we collect remarks on regularity conditions and some additional preliminaries.

### B.1 Additional Remarks on Informal Assumption 1.1 and Assumption 3.1

*Remark* B.1 (Thresholds). In the theorem below we take $\tau_a = 96KT(a)$ for $a \in \{0, 1\}$. Any threshold of the same order gives the same bound after changing constants. The truncation by $m^{-1}$ in the definition of $\widehat{p}_i(a)$ is only to avoid division by zero; under Assumption 3.1, it is inactive with high probability.

*Remark* B.2 (Why the empirical rank is well-defined). The signal-to-noise condition ensures that the gap-selection step in Algorithm 1 is nonempty with high probability. Since $\mathrm{rank}(A(a)) \le r$, we have $\sigma_1(a) = \sum_{\ell=1}^{r}(\sigma_\ell(a) - \sigma_{\ell+1}(a))$, with $\sigma_{r+1}(a) = 0$. Hence $\sigma_1(a) > 120KrT(a)$ implies that

$\sigma_s(a) - \sigma_{s+1}(a) > 120KT(a)$ for some $s \le r$. On the high-probability event $\|\widetilde{Y}^{\mathrm{ub}}(a) - A(a)\|_{\mathrm{op}} \le 12KT(a)$, Weyl's inequality gives $\widetilde{\sigma}_s^{\mathrm{ub}}(a) - \widetilde{\sigma}_{s+1}^{\mathrm{ub}}(a) \ge 96KT(a)$, so the algorithm selects at least one admissible truncation level. The selected $\widehat{s}(a)$ is therefore an empirical effective rank, not necessarily the algebraic rank of $A(a)$.

### B.2 Some Classical Results

In this section, we recall standard results used in Section 3, Section 4, and Section C.1.

**Theorem B.3 (Weyl's inequality** Weyl (1912)**).** *Let $A, E$ be $n \times m$ matrices, and define $\widetilde{A} := A + E$. Then, for any $1 \le i \le \min\{m, n\}$,*

$$|\widetilde{\lambda}_i - \lambda_i| \le \|E\|_{\mathrm{op}} \quad \textit{and} \quad |\widetilde{\sigma}_i - \sigma_i| \le \|E\|_{\mathrm{op}}\,,$$

*where $\lambda_i, \widetilde{\lambda}_i$ are the $i$th eigenvalues of $A$ and $\widetilde{A}$, and $\sigma_i, \widetilde{\sigma}_i$ are the corresponding singular values.*

**Theorem B.4 (Cauchy's integral theorem).** *Let $\Gamma$ be a simple closed contour, and let $f$ be an analytic function in the whole simply connected domain $S$ containing $\Gamma$. Then*

$$\frac{1}{2\pi \mathbf{i}} \int_\Gamma \frac{f(z)}{z-a}\,\mathrm{d}z = \begin{cases} f(a), & a \textit{ inside } \Gamma\,, \\ 0, & a \textit{ outside } \Gamma \end{cases}.$$

*In particular, for $f(z) = z$, one has*

$$\frac{1}{2\pi \mathbf{i}} \int_\Gamma \frac{z}{z-a}\,\mathrm{d}z = \begin{cases} a, & a \textit{ inside } \Gamma\,, \\ 0, & a \textit{ outside } \Gamma \end{cases}.$$

**Theorem B.5 (Eckart–Young–Mirsky bound** Eckart & Young (1936)**).** *Let $A, \widetilde{A} \in \mathbb{R}^{n\times m}$, and let $A_s$, $\widetilde{A}_s$ denote their respective best rank-$s$ approximations. Set $E := \widetilde{A} - A$. Then,*

$$\|\widetilde{A}_s - A_s\|_{\mathrm{op}} \le 2\,(\sigma_{s+1} + \|E\|_{\mathrm{op}})\,,$$

*where $\sigma_{s+1}$ is the $(s+1)$th singular value of $A$.*

## C Perturbation of Low-Rank Approximations in Row-Wise $\ell_2$ Norm

We recall the definition of the row-wise $\ell_2$ norm. For each natural number $N$, denote the standard basis of $\mathbb{R}^N$ by $\{e_{N,k} \text{ for } 1 \le k \le N\}$. For an $n \times m$ matrix $M$ with rows $r_1, r_2, \cdots, r_n$, define

$$\|M\|_{2,\infty} := \max_{1\le i\le n} \|r_i\|_2 = \max_{1\le i\le n} \|e_{n,i}^\top M\|_2\,.$$

We return to the perturbation of low-rank approximations. Let $A$ be an $n \times m$ matrix with singular value decomposition

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top\,,$$

where $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r$ are singular values with the corresponding pairs of left and right singular vectors $(u_i, v_i)_{1\le i\le r}$. For each $1 \le i \le r-1$, the $i$th singular gap is $\delta_i := \sigma_i - \sigma_{i+1}$.

Let $E_R$ be an $n \times m$ random matrix of mean zero and $E_0$ be an $n \times m$ deterministic matrix. We denote

$$E = E_R + E_0 \qquad \text{and} \qquad \widetilde{A} = A + E\,.$$

For a given rank-parameter $s$, our goal is to bound

$$\|\widetilde{A}_s - A_s\|_{2,\infty}\,.$$

Before stating our main bound, we recall the following notions and definitions.

**Incoherence number.** Define

$$\mu_R := \sqrt{n} \max_{1\le i\le n} \|e_{n,i}^\top U\|_2 \qquad \text{and} \qquad \mu_C := \sqrt{m} \cdot \max_{1\le j\le m} \|V e_{m,j}\|_2\,.$$

The *incoherence number* of $A$ is $\mu := \max\{\mu_R, \mu_C\}$.

**Definition C.1.** A random matrix $E_R$ is called $(K, \sigma)$-bounded if $E_R$ has entries satisfying
$\mathbb{E}\, E_{R,ij} = 0$, $\mathbb{E}[|E_{R,ij}|^2] \le \sigma^2$, $\mathbb{E}[|E_{R,ij}|^l] \le K^{l-2}\sigma^2$ for all $l \in \mathbb{N}, l \ge 2$, and $i \in [n], j \in [m]$.

We restate our key result - Theorem 4.2:

**Theorem 4.2.** Let $E_R$ be a $(K, \sigma)$-bounded random matrix. There is a universal constant $C > 0$ such that the following holds. If $\delta_s \ge 6(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}})$, then with probability at least $1 - O(1/(m+n))$,

$$\frac{\|\widetilde{A}_s - A_s\|_{2,\infty}}{C\sqrt{r}} \le \left[\log^2(m+n) + \frac{K\log^4(m+n)}{\sqrt{m+n}}\right]\left(\frac{\mu}{\sqrt{m}} + \frac{\mu}{\sqrt{n}}\right)\cdot\sqrt{m+n}\,\frac{\sigma_s}{\delta_s}\left[\sigma + \|E_0\|_{\mathrm{op}}\right]. \quad \text{(C.1)}$$

To simplify the presentation, we focus on the case $n = \Theta(m)$, in which the RHS simplifies to

$$\left[\log^2 m + \frac{K\log^4 m}{\sqrt{m}}\right](\sigma + \|E_0\|_{\mathrm{op}})\,\mu\frac{\sigma_s}{\delta_s}.$$

We obtain the following corollaries in this setting. The first one is when $E_0 = 0$ (purely random noise), and the second one is when $\|E_0\|_{\mathrm{op}}$ is large.

**Corollary C.2** (Random perturbation)**.** *Under the above setting, let $E_R$ be a $(K, \sigma)$-bounded random matrix. There is a universal constant $C > 0$ such that the following holds. If $\delta_s \ge 6\|E_R\|_{\mathrm{op}}$, then with probability at least $1 - O(m^{-1})$,*

$$\|(A + E_R)_s - A_s\|_{2,\infty} \le C\sqrt{r}\left[\log^2 m + \frac{K\log^4 m}{\sqrt{m}}\right]\mu\sigma\frac{\sigma_s}{\delta_s}. \quad \text{(C.2)}$$

If $r, K, \sigma, \mu$ are all $\widetilde{O}(1)$, then the RHS becomes $\widetilde{O}\big(\frac{\sigma_s}{\delta_s}\big)$. Moreover, in many widely studied structured models (*e.g.*, spiked covariance, stochastic block, and graph Laplacian models), one typically has $\sigma_s = O(\delta_s)$, yielding the clean bound $\widetilde{O}(1)$. Compared with existing bounds on $\|(A+E_R)_s - A_s\|_{\mathrm{op}}$ *e.g.*, Tran & Vu (2025c); Tran et al. (2025); Mangoubi & Vishnoi (2025); Eckart & Young (1936), which are at best $O(\|E_R\|_{\mathrm{op}}) = O(\sqrt{m})$, our $\|\cdot\|_{2,\infty}$-bound is smaller by a factor of $\sqrt{m}$. Since the optimal bound for $\|(A + E_R)_s - A_s\|_{\mathrm{op}}$ is $O(\|E_R\|_{\mathrm{op}})$, and

$$\|(A + E_R)_s - A_s\|_{\mathrm{op}} \le \sqrt{n}\|(A + E_R)_s - A_s\|_{2,\infty} = \Theta(\sqrt{m}\|(A + E_R)_s - A_s\|_{2,\infty}),$$

our bound is sharp up to a logarithmic factor.

**Corollary C.3.** *In the above setting, let $E_R$ be a $(K, \sigma)$-bounded random matrix and $\|E_0\|_{\mathrm{op}} \gg \sigma$. There is a universal constant $C > 0$ such that the following holds. If $\delta_s \ge 6(\|E_R\|_{\mathrm{op}} + \|E_0\|_{\mathrm{op}})$, then with probability at least $1 - O(m^{-1})$,*

$$\|(A + E_R + E_0)_s - A_s\|_{2,\infty} \le C\sqrt{r}\left[\log^2 m + \frac{K\log^4 m}{\sqrt{m}}\right]\mu\frac{\sigma_s}{\delta_s}\|E_0\|_{\mathrm{op}}. \quad \text{(C.3)}$$

Similar to the discussion after Corollary C.2, in many typical settings the right-hand side simplifies to $\widetilde{O}(\|E_0\|_{\mathrm{op}})$, which is sharp up to logarithmic factors. For example, consider $m = n$, $E_R = 0$, $A$ is a diagonal matrix, and $E_0 = cI_m$. Then

$$\|(A + E_R + E_0)_s - A_s\|_{2,\infty} = \sigma_1 + c - \sigma_1 = c = \|E_0\|_{\mathrm{op}}.$$

Unlike random noise $E_R$, whose effect spreads across all directions, the deterministic component $E_0$ can have a localized impact on the perturbation.

### C.1 Proof of Theorem 4.2

In this subsection, we prove Theorem 4.2. Recall that for a vector $\mathbf{u} \in \mathbb{R}^N$, $\|\mathbf{u}\|_\infty := \max_{1\le i\le N}|\mathbf{u}_i|$.

**Step 1 - Symmetrization.** We symmetrize $A$ and $\widetilde{A}$ as follows. Set $\mathcal{A} := \begin{pmatrix} 0 & A \\ A^\top & 0 \end{pmatrix}$, $\mathcal{E} := \begin{pmatrix} 0 & E \\ E^\top & 0 \end{pmatrix}$, and then $\widetilde{\mathcal{A}} = \begin{pmatrix} 0 & \widetilde{A} \\ \widetilde{A}^\top & 0 \end{pmatrix}$. Given the singular value decomposition $A = U\Sigma V^\top$, it is easy to see that $\mathcal{A}$ admits the following spectral decomposition

$$\mathcal{A} = \begin{pmatrix} \frac{U}{\sqrt{2}} & \frac{U}{\sqrt{2}} \\ \frac{V}{\sqrt{2}} & -\frac{V}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{pmatrix}\begin{pmatrix} \frac{U^\top}{\sqrt{2}} & \frac{V^\top}{\sqrt{2}} \\ \frac{U^\top}{\sqrt{2}} & -\frac{V^\top}{\sqrt{2}} \end{pmatrix},$$

that is, $\mathcal{A}$ has the eigenvalues $\pm\sigma_i$ with the corresponding eigenvector $\mathbf{u}_{\pm i} \coloneqq \frac{1}{\sqrt{2}}\begin{pmatrix} u_i \\ \pm v_i \end{pmatrix}$. Similarly, $\widetilde{\mathcal{A}}$ has the eigenvalues $\pm\widetilde{\sigma}_i$ with the corresponding eigenvector $\widetilde{\mathbf{u}}_{\pm i}$.

The best rank-$2s$ approximation of $\mathcal{A}$ is

$$\mathcal{A}_{2s} \coloneqq \begin{pmatrix} \frac{U}{\sqrt{2}} & \frac{U}{\sqrt{2}} \\ \frac{V}{\sqrt{2}} & -\frac{V}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \Sigma_s & 0 \\ 0 & -\Sigma_s \end{pmatrix} \begin{pmatrix} \frac{U^\top}{\sqrt{2}} & \frac{V^\top}{\sqrt{2}} \\ \frac{U^\top}{\sqrt{2}} & -\frac{V^\top}{\sqrt{2}} \end{pmatrix},$$

which, in fact, equals

$$\mathcal{A}_{2s} = \begin{pmatrix} 0 & U\Sigma_s V^\top \\ V\Sigma_s U^\top & 0 \end{pmatrix}.$$

Similarly,

$$\widetilde{\mathcal{A}}_{2s} = \begin{pmatrix} 0 & \widetilde{U}\widetilde{\Sigma}_s\widetilde{V}^\top \\ \widetilde{V}\widetilde{\Sigma}_s\widetilde{U}^\top & 0 \end{pmatrix},$$

and then,

$$\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s} = \begin{pmatrix} 0 & \widetilde{U}\widetilde{\Sigma}_s\widetilde{V}^\top - U\Sigma_s V^\top \\ \widetilde{V}\widetilde{\Sigma}_s\widetilde{U}^\top - V\Sigma_s U^\top & 0 \end{pmatrix}.$$

This gives

$$\left\|e_1^\top(\widetilde{A}_s - A_s)\right\|_2 = \left\|e_1^\top(\widetilde{U}\widetilde{\Sigma}_s\widetilde{V}^\top - U\Sigma_s V^\top)\right\|_2 = \left\|e_1^\top(\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s})\right\|_2,$$

and then

$$\|\widetilde{A}_s - A_s\|_{2,\infty} \coloneqq \max_{1\le k\le n} \|e_{n,k}^\top(\widetilde{A}_s - A_s)\|_2 = \max_{1\le k\le n} \left\|e_{m+n,k}^\top(\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s})\right\|_2. \tag{C.4}$$

**Step 2 - Contour representation of perturbation.** Let $\Gamma$ be a contour in $\mathbb{C}$ that encloses $\pm\sigma_1, \pm\sigma_2, \ldots, \pm\sigma_s$ and excludes $\pm\sigma_{s+1}, \pm\sigma_{s+2}, \ldots, \pm\sigma_r$. The well-known contour integral representation Higham (2008); Kato (1980); Stewart & Sun (1990) gives us:

$$\frac{1}{2\pi\mathbf{i}}\int_\Gamma z(zI - \mathcal{A})^{-1}\,\mathrm{d}z = \sum_{i=1}^s \sigma_i \mathbf{u}_{+i}\mathbf{u}_{+i}^\top + \sum_{i=1}^s (-\sigma_i)\mathbf{u}_{-i}\mathbf{u}_{-i}^\top = \mathcal{A}_{2s}.$$

Let $\widetilde{\lambda}_1 \ge \cdots \ge \widetilde{\lambda}_{m+n}$ denote the eigenvalues of $\widetilde{\mathcal{A}}$. The construction of $\Gamma$ (presented later) and the gap assumption $6\|E\|_{\mathrm{op}} < \delta_s$ ensure that the eigenvalues $\widetilde{\lambda}_i$ lie inside $\Gamma$ if and only if $|\widetilde{\lambda}_i| \in \{\widetilde{\sigma}_1, \widetilde{\sigma}_2, \ldots, \widetilde{\sigma}_s\}$. Then, similarly, we have $\frac{1}{2\pi\mathbf{i}}\int_\Gamma z(zI - \widetilde{\mathcal{A}})^{-1}\,\mathrm{d}z = \sum_{i=1}^s \widetilde{\sigma}_i\widetilde{\mathbf{u}}_{+i}\widetilde{\mathbf{u}}_{+i}^\top + \sum_{i=1}^s(-\widetilde{\sigma}_i)\widetilde{\mathbf{u}}_{-i}\widetilde{\mathbf{u}}_{-i}^\top = \widetilde{\mathcal{A}}_{2s}$.

Thus, we obtain the following contour identity for the perturbation:

$$\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s} = \frac{1}{2\pi\mathbf{i}}\int_\Gamma z[(zI - \widetilde{\mathcal{A}})^{-1} - (zI - \mathcal{A})^{-1}]\,\mathrm{d}z. \tag{C.5}$$

Therefore,

$$e_1^\top\left(\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s}\right) = \frac{1}{2\pi\mathbf{i}}\int_\Gamma z \cdot e_1^\top[(zI - \widetilde{\mathcal{A}})^{-1} - (zI - \mathcal{A})^{-1}]\,\mathrm{d}z.$$

This contour method is widely used in functional perturbation analysis, *e.g.*, Higham (2008); Kato (1980); Koltchinskii & Xia (2016); O'Rourke et al. (2023); Tran & Vu (2025b); Tran et al. (2025); Tran & Vu (2026a); Tran & Vishnoi (2025). Note that one can trivially bound $\|e_1^\top\left(\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s}\right)\|_2$ by $\|\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s}\|_{\mathrm{op}}$, and then apply the existing bounds (*e.g.*, Tran et al. (2025); Tran & Vu (2025c); Eckart & Young (1936)) in perturbation theory. However, this approach only gives a suboptimal bound that is off by a factor of up to $\sqrt{n+m}$. Thus, obtaining a sharp bound with respect to the $\|\cdot\|_{2,\infty}$-norm remains a formidable analytical challenge.

**Step 3 - Contour expansion.** We adapt the contour expansion method, introduced in Tran & Vu (2026a; 2025a); Tran et al. (2025). Using the Sherman-Morrison-Woodbury formula $M^{-1} - (M+N)^{-1} = (M+N)^{-1}NM^{-1}$ (Horn & Johnson, 2012) and the fact that $\widetilde{\mathcal{A}} = \mathcal{A} + \mathcal{E}$, we obtain

$$(zI - \mathcal{A})^{-1} - (zI - \widetilde{\mathcal{A}})^{-1} = (zI - \mathcal{A})^{-1}\mathcal{E}(zI - \widetilde{\mathcal{A}})^{-1},$$

and hence

$$(zI - \mathcal{A})^{-1} - (zI - \widetilde{\mathcal{A}})^{-1} = (zI - \mathcal{A})^{-1}\mathcal{E}(zI - \mathcal{A})^{-1} + (zI - \mathcal{A})^{-1}\mathcal{E}[(zI - \widetilde{\mathcal{A}})^{-1} - (zI - \mathcal{A})^{-1}].$$

By repeatedly applying the resolvent formula $L$ times, one can expand $ze_1^\top[(zI - \widetilde{\mathcal{A}})^{-1} - (zI - \mathcal{A})^{-1}]$ into

$$\left(\sum_{k=1}^{L} z \cdot e_1^\top (zI - \mathcal{A})^{-1}[\mathcal{E}(zI - \mathcal{A})^{-1}]^k\right) + e_1^\top (zI - \mathcal{A})^{-1}[\mathcal{E}(zI - \mathcal{A})^{-1}]^{L-1} \cdot [(zI - \widetilde{\mathcal{A}})^{-1} - (zI - \mathcal{A})^{-1}].$$

Therefore,

$$\begin{aligned} &e_1^\top \left(\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s}\right) = \left(\sum_{k=1}^{L} F_k\right) + F_{\text{tail}}, \quad \text{where} \\ &F_k \coloneqq \frac{1}{2\pi \mathbf{i}} \int_\Gamma z \cdot e_1^\top (zI - \mathcal{A})^{-1}[\mathcal{E}(zI - \mathcal{A})^{-1}]^k \, \mathrm{d}z \quad \text{and} \\ &F_{\text{tail}} \coloneqq \frac{1}{2\pi \mathbf{i}} \int_\Gamma e_1^\top (zI - \mathcal{A})^{-1}[\mathcal{E}(zI - \mathcal{A})^{-1}]^{L-1} \cdot [(zI - \widetilde{\mathcal{A}})^{-1} - (zI - \mathcal{A})^{-1}] \, \mathrm{d}z. \end{aligned} \tag{C.6}$$

Thus, by the triangle inequality, we have

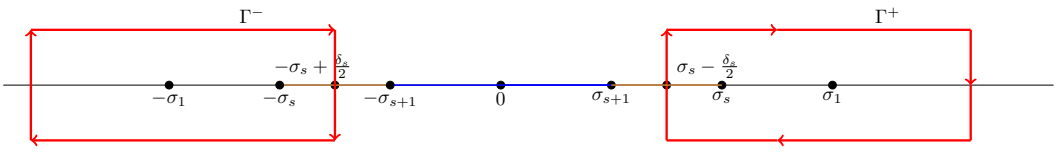
$$\left\| e_1^\top \left(\widetilde{\mathcal{A}}_{2s} - \mathcal{A}_{2s}\right) \right\|_2 \le \left(\sum_{k=1}^{L} \|F_k\|_2\right) + \|F_{\text{tail}}\|_2.$$

We set $L = (\ell + 1)\log(m+n)$ (the constant $\ell$ will be chosen later). It remains to bound $\|F_k\|_2$ for each $1 \le k \le L$ and $\|F_{\text{tail}}\|_2$. We will show that the dominant contribution is $\sum_{k=1}^{L} \|F_k\|_2$ and $\|F_{\text{tail}}\|_2$ is negligible.

Inspired by the construction of $\Gamma$ in Tran & Vu (2025b); Tran et al. (2025); Tran & Vu (2026b), we define $\Gamma$ as the union of two disjoint rectangles $\Gamma^+ \cup \Gamma^-$, where:

- $\Gamma^+$ isolates $\{\sigma_1, \sigma_2, \ldots, \sigma_s\}$, with (1) its left vertical edge intersecting the real axis at $a_0 \coloneqq \sigma_s - \delta_s/2$; (2) its right vertical edge intersecting the real axis at $a_1 \coloneqq 2\sigma_1$; (3) its height (from the real line) is $T = 2\sigma_1$.
- $\Gamma^-$ isolates $\{-\sigma_1, -\sigma_2, \ldots, -\sigma_s\}$, with (1) its left vertical edge intersecting the real axis at $b_1 \coloneqq -2\sigma_1$; (2) its right vertical edge intersecting the real axis at $b_0 \coloneqq -\sigma_s + \delta_s/2$; (3) its height (from the real line) is $T = 2\sigma_1$.

See the figure below for an illustration.




**Step 4 - Bounding $\|F_k\|$ for a natural number $k \le L$.** We split the spectral decomposition of $(zI - \mathcal{A})^{-1} = \sum_{i=1}^{r} \frac{\mathbf{u}_{+i}\mathbf{u}_{+i}^\top}{z - \sigma_i} + \sum_{i=1}^{r} \frac{\mathbf{u}_{-i}\mathbf{u}_{-i}^\top}{z + \sigma_i} + \frac{I_{m+n} - (\sum_{i=1}^{r} \mathbf{u}_{+i}\mathbf{u}_{+i}^\top + \mathbf{u}_{-i}\mathbf{u}_{-i}^\top)}{z}$ into $P + Q$, where

$$\begin{aligned} P &\coloneqq \sum_{i=1}^{r} \frac{\mathbf{u}_{+i}\mathbf{u}_{+i}^\top}{z - \sigma_i} + \sum_{i=1}^{r} \frac{\mathbf{u}_{-i}\mathbf{u}_{-i}^\top}{z + \sigma_i} - \frac{\sum_{i=1}^{r} \mathbf{u}_{+i}\mathbf{u}_{+i}^\top + \mathbf{u}_{-i}\mathbf{u}_{-i}^\top}{z} \\ &= \sum_{i=1}^{r} \frac{\sigma_i}{(z - \sigma_i)z} \mathbf{u}_{+i}\mathbf{u}_{+i}^\top + \sum_{i=1}^{r} \frac{-\sigma_i}{(z + \sigma_i)z} \mathbf{u}_{-i}\mathbf{u}_{-i}^\top, \\ \text{and} \quad Q &\coloneqq \frac{I_{m+n}}{z}. \end{aligned} \tag{C.7}$$

We can rewrite $2\pi\mathbf{i}F_k$ as follows:

$$\begin{aligned}
&\int_\Gamma z \cdot e_1^\top P[\mathcal{E}(zI-\mathcal{A})^{-1}]^k \,\mathrm{d}z + \int_\Gamma z \cdot e_1^\top Q[\mathcal{E}(zI-\mathcal{A})^{-1}]^k \,\mathrm{d}z \\
&= \int_\Gamma z \cdot e_1^\top P[\mathcal{E}(zI-\mathcal{A})^{-1}]^k \,\mathrm{d}z + \int_\Gamma z \cdot \frac{1}{z} e_1^\top \mathcal{E}(zI-\mathcal{A})^{-1} \cdot [\mathcal{E}(zI-\mathcal{A})^{-1}]^{k-1} \,\mathrm{d}z \\
&= \int_\Gamma z \cdot e_1^\top P[\mathcal{E}(zI-\mathcal{A})^{-1}]^k \,\mathrm{d}z + \int_\Gamma z \cdot \frac{1}{z} e_1^\top \mathcal{E}P \cdot [\mathcal{E}(zI-\mathcal{A})^{-1}]^{k-1} \,\mathrm{d}z \\
&\quad + \int_\Gamma z \cdot \frac{1}{z^2} e_1^\top \mathcal{E} \cdot [\mathcal{E}(zI-\mathcal{A})^{-1}]^{k-1} \,\mathrm{d}z \\
&= \cdots \\
&= \left[\sum_{l=0}^{k} \int_\Gamma z \cdot \frac{1}{z^l} e_1^\top \mathcal{E}^l P[\mathcal{E}(zI-\mathcal{A})^{-1}]^{k-l} \,\mathrm{d}z\right] + \int_\Gamma z \cdot \frac{1}{z^{k+1}} e_1^\top \mathcal{E}^k \,\mathrm{d}z\,.
\end{aligned} \tag{C.8}$$

Note that $z = 0$ is outside of the contour $\Gamma$. Thus, the last term $\int_\Gamma z \cdot \frac{1}{z^{k+1}} e_1^\top \mathcal{E}^k \,\mathrm{d}z$ is zero, and hence

$$F_k \coloneqq \sum_{l=0}^{k} \frac{1}{2\pi\mathbf{i}} \int_\Gamma z \cdot \frac{1}{z^l} e_1^\top \mathcal{E}^l P[\mathcal{E}(zI-\mathcal{A})^{-1}]^{k-l} \,\mathrm{d}z\,.$$

By the triangle inequality, we have

$$\|F_k\|_2 \le \sum_{l=0}^{k} G_{k,l} \quad \text{where} \quad G_{k,l} \coloneqq \frac{1}{2\pi} \cdot \int_\Gamma \left\| \frac{1}{z^{l-1}} e_1^\top \mathcal{E}^l P[\mathcal{E}(zI-\mathcal{A})^{-1}]^{k-l} \right\|_2 |\,\mathrm{d}z|\,.$$

If $l \ge 1$, splitting out the factor $[\mathcal{E}(zI-\mathcal{A})^{-1}]^{k-l}$, we have

$$\begin{aligned}
G_{k,l} &\le \frac{\max_{z\in\Gamma} \|\mathcal{E}(zI-\mathcal{A})^{-1}\|_{\mathrm{op}}^{k-l}}{2\pi} \cdot \int_\Gamma \left\| \frac{1}{z^{l-1}} e_1^\top \mathcal{E}^l P \right\|_2 |\,\mathrm{d}z| \\
&\le \frac{1}{2\cdot 3^{k-l}\pi} \cdot \int_\Gamma \left\| \frac{1}{z^{l-1}} e_1^\top \mathcal{E}^l P \right\|_2 |\,\mathrm{d}z|\,.
\end{aligned}$$

The last inequality is true by our gap assumption that $6\|\mathcal{E}\|_{\mathrm{op}} < \delta_s$ and the construction of $\Gamma$, which implies $\min_{z\in\Gamma, i\in[m+n]} |z-\lambda_i| \ge \delta_s/2 \ge 3\|\mathcal{E}\|_{\mathrm{op}}$, and hence

$$\max_{z\in\Gamma} \|\mathcal{E}(zI-\mathcal{A})^{-1}\|_{\mathrm{op}} \le \max_{z\in\Gamma} \|(zI-\mathcal{A})^{-1}\|_{\mathrm{op}} \cdot \|\mathcal{E}\|_{\mathrm{op}} = \frac{\|\mathcal{E}\|_{\mathrm{op}}}{\min_{z\in\Gamma, i\in[n]} |z \pm \sigma_i|} \le \frac{\|\mathcal{E}\|_{\mathrm{op}}}{3\|\mathcal{E}\|_{\mathrm{op}}} = \frac{1}{3}\,.$$

For the second factor, we use the following lemma. Its proof appears in the next section.

**Lemma C.4.** *Fix a natural number $l \ge 1$. Under the assumptions of Theorem 4.2, with probability at least $1 - O((m+n)^{-2})$, there is a universal constant $C > 0$ such that $\frac{\int_\Gamma \left\| \frac{1}{z^{l-1}} e_1^\top \mathcal{E}^l P \right\|_2 |\,\mathrm{d}z|}{2\pi C\sqrt{r}}$ is at most*

$$\begin{aligned}
&\frac{1}{3^{l-1}} \cdot (\sigma\sqrt{m+n}\frac{\sigma_s}{\delta_s}) \\
&\quad \times \left( \left[\log^2(m+n) + \frac{K\log^4(m+n)}{\sqrt{m+n}}\right] \max_{1\le i\le r} \|\mathbf{u}_{\pm i}\|_\infty + \frac{\|\mathcal{E}_0\|_{\mathrm{op}} \log(m+n)}{\sigma\sqrt{m+n}} \right)\,.
\end{aligned}$$

Thus, by Lemma C.4, we further obtain that $G_{k,l}, l \ge 1$ is at most

$$\begin{aligned}
&\frac{C\sqrt{r}}{3^{k-1}} \cdot (\sigma\sqrt{m+n}\frac{\sigma_s}{\delta_s}) \\
&\quad \times \left( \left[\log^2(m+n) + \frac{K\log^4(m+n)}{\sqrt{m+n}}\right] \max_{1\le i\le r} \|\mathbf{u}_{\pm i}\|_\infty + \frac{\|\mathcal{E}_0\|_{\mathrm{op}} \log(m+n)}{\sigma\sqrt{m+n}} \right)\,,
\end{aligned} \tag{C.9}$$

with probability at least $1 - O((m+n)^{-2})$ for some universal constant $C$.

Next, for $l = 0$, splitting out the factor $[\mathcal{E}(zI - \mathcal{A})^{-1}]^{k-1}$, we have

$$\begin{aligned} G_{k,0} &\le \frac{\max_{z\in\Gamma} \|\mathcal{E}(zI-\mathcal{A})^{-1}\|_{\mathrm{op}}^{k-1}}{2\pi} \cdot \int_\Gamma \left\| z e_1^\top P\mathcal{E}(zI-\mathcal{A})^{-1} \right\|_2 |\,\mathrm{d}z| \\ &\le \frac{1}{2\cdot 3^{k-1}\pi} \cdot \int_\Gamma \left\| z e_1^\top P\mathcal{E}(zI-\mathcal{A})^{-1} \right\|_2 |\,\mathrm{d}z|\,. \end{aligned}$$

Similarly, for the second factor, we use the following technical lemma, whose proof will be presented in the next section.

**Lemma C.5.** *Under the assumptions of Theorem 4.2, with probability at least* $1 - O((m+n)^{-2})$*, there is a universal constant* $C > 0$ *such that*

$$\frac{\int_\Gamma \left\| z e_1^\top P\mathcal{E}(zI-\mathcal{A})^{-1} \right\|_2 |\,\mathrm{d}z|}{2\pi} \le C\sqrt{r}\log^2(m+n)\cdot(\|\mathcal{E}_0\|_{\mathrm{op}}+\sigma)(\sqrt{m+n}\frac{\sigma_s}{\delta_s})\cdot \max_{1\le i\le r}\|\mathbf{u}_{\pm i}\|_\infty\,.$$

By Lemma C.5, with probability at least $1 - O((m+n)^{-2})$,

$$\frac{\int_\Gamma \left\| z e_1^\top P\mathcal{E}(zI-\mathcal{A})^{-1} \right\|_2 |\,\mathrm{d}z|}{2\pi} \le C\sqrt{r}\log^2(m+n)\cdot(\|\mathcal{E}_0\|_{\mathrm{op}}+\sigma)(\sqrt{m+n}\frac{\sigma_s}{\delta_s})\cdot \max_{1\le i\le r}\|\mathbf{u}_{\pm i}\|_\infty\,.$$

Thus,

$$G_{k,0} \le \frac{C\sqrt{r}}{3^{k-1}}\log^2(m+n)\cdot(\|\mathcal{E}_0\|_{\mathrm{op}}+\sigma)\cdot(\sqrt{m+n}\frac{\sigma_s}{\delta_s})\cdot \max_{1\le i\le r}\|\mathbf{u}_{\pm i}\|_\infty\,. \tag{C.10}$$

The estimates on $G_{k,l}$ and the fact that $\max_{1\le i\le r}\|\mathbf{u}_{\pm i}\|_\infty \ge \frac{1}{\sqrt{m+n}}$ imply that with probability at least $1 - O(k(m+n)^{-2})$, there is a universal constant $C' > 3C$, such that

$$\begin{aligned} \|F_k\| &\le \sum_{l=0}^{k} G_{k,l} \\ &\le \frac{C'k\sqrt{r}}{3^k}\cdot(\|\mathcal{E}_0\|_{\mathrm{op}}+\sigma)(\sqrt{m+n}\frac{\sigma_s}{\delta_s})\left[\log^2(m+n)+\frac{K\log^4(m+n)}{\sqrt{m+n}}\right]\max_{1\le i\le r}\|\mathbf{u}_{\pm i}\|_\infty\,. \end{aligned} \tag{C.11}$$

**Step 5 - Bounding $\|F_{\mathrm{tail}}\|$.** Repeating the contour expansion in Step 3, we can decompose $F_{\mathrm{tail}}$ into

$$\sum_{s>L}\frac{1}{2\pi\mathbf{i}}\int_\Gamma e_1^\top(zI-\mathcal{A})^{-1}[\mathcal{E}(zI-\mathcal{A})^{-1}]^s dz.$$

Thus, by the triangle inequality, we have

$$\begin{aligned} \|F_{\mathrm{tail}}\|_2 &\le \frac{1}{2\pi}\sum_{s>L}\Big\|\int_\Gamma e_1^\top(zI-\mathcal{A})^{-1}[\mathcal{E}(zI-\mathcal{A})^{-1}]^s dz\Big\|_2 \\ &\le \frac{1}{2\pi}\sum_{s>L}\int_\Gamma \left\|e_1^\top(zI-\mathcal{A})^{-1}[\mathcal{E}(zI-\mathcal{A})^{-1}]^s\right\|_2 |dz|. \end{aligned}$$

Since the length of $\Gamma$ is at most $24\sigma_1$, we further obtain

$$\begin{aligned} \|F_{\mathrm{tail}}\|_2 &\le \frac{24\sigma_1}{2\pi}\sum_{s>L}\max_{z\in\Gamma}\|e_1^\top(zI-\mathcal{A})^{-1}[\mathcal{E}(zI-\mathcal{A})^{-1}]^s\|_2 \\ &\le \frac{24\sigma_1}{2\pi}\sum_{s>L}\max_{z\in\Gamma}\|(zI-\mathcal{A})^{-1}\|_{\mathrm{op}}\cdot[\max_{z\in\Gamma}\|\mathcal{E}(zI-\mathcal{A})^{-1}\|_{\mathrm{op}}]^s. \end{aligned} \tag{C.12}$$

Moreover, together, the construction of $\Gamma$ and the assumption that $\delta_s \ge 6\|\mathcal{E}\|_{\mathrm{op}}$ imply

$$\begin{aligned} &\max_{z\in\Gamma}\|(zI-\mathcal{A})^{-1}\|_{\mathrm{op}} = \frac{1}{\min_j|z-\lambda_j|} = \frac{1}{\delta_s/2} = \frac{2}{\delta_s}, \\ &\max_{z\in\Gamma}\|\mathcal{E}(zI-\mathcal{A})^{-1}\|_{\mathrm{op}} \le \|\mathcal{E}\|_{\mathrm{op}}\cdot\max_{z\in\Gamma}\|(zI-\mathcal{A})^{-1}\|_{\mathrm{op}} = \frac{2\|\mathcal{E}\|_{\mathrm{op}}}{\delta_s} \le \frac{1}{3}. \end{aligned} \tag{C.13}$$

Substituting (C.13) into (C.12), we obtain

$$\begin{aligned}\|F_{\text{tail}}\|_2 &\le \frac{24\sigma_1}{2\pi}\sum_{s>L}\frac{2}{\delta_s}\cdot\frac{1}{3^s}\\ &= \frac{24\sigma_1}{\pi\delta_s}\cdot\frac{1}{3^L}\\ &= \left(\frac{\|\mathcal{E}\|_{\text{op}}\cdot\sigma_s}{\delta_s}\cdot\frac{1}{\sqrt{m+n}}\right)\cdot\frac{24\sigma_1\sqrt{m+n}}{\pi\cdot 3^L\|\mathcal{E}\|_{\text{op}}\cdot\sigma_s}\end{aligned} \tag{C.14}$$

Let $\ell$ be the natural number such that

$$\frac{\sigma_1}{\sigma_s\|\mathcal{E}\|_{\text{op}}} \le (m+n)^{\ell}.$$

Thus, by setting $L=(\ell+1)\log(m+n)$ in (C.14), we have

$$\|F_{\text{tail}}\|_2 = o\left(\frac{\|\mathcal{E}\|_{\text{op}}\cdot\sigma_s}{\delta_s}\cdot\frac{1}{\sqrt{m+n}}\right),$$

which is negligible in comparison to $\sum_{k=1}^{L}\|F_k\|_2$.

Combining the estimates on $\|F_k\|_2$ and $\|F_{\text{tail}}\|_2$, we obtain

$$\begin{aligned}\left\|e_1^\top\left(\widetilde{\mathcal{A}}_{2s}-\mathcal{A}_{2s}\right)\right\|_2 \le (C'+o(1))\sqrt{r}\cdot(\|\mathcal{E}_0\|_{\text{op}}+\sigma)(\sqrt{m+n}\frac{\sigma_s}{\delta_s})\\ \times\left[\log^2(m+n)+\frac{K\log^4(m+n)}{\sqrt{m+n}}\right]\max_{1\le i\le r}\|\mathbf{u}_{\pm i}\|_\infty.\end{aligned} \tag{C.15}$$

Since $\mathbf{u}_{\pm i}\coloneqq\frac{1}{\sqrt{2}}\begin{pmatrix}u_i\\ \pm v_i\end{pmatrix}$, by the definition of the incoherence parameter $\mu$, we have

$$\max_{1\le i\le r}\|\mathbf{u}_{\pm i}\|_\infty \le \frac{\mu}{\sqrt{m}}+\frac{\mu}{\sqrt{n}}.$$

Combining (C.4) with (C.15) and this upper bound of $\max_{1\le i\le r}\|\mathbf{u}_{\pm i}\|_\infty$, we prove Theorem 4.2.

*Remark* C.6. We can slightly improve the upper bound on $\|F_k\|_2$ by combining exactly the upper bounds on $G_{k,0}$ and $G_{k,l}$ without replacing $\frac{1}{\sqrt{m+n}}$ by $\max_{1\le i\le r}\|\mathbf{u}_{\pm i}\|_\infty$. Indeed, $\|F_k\|_2/(C\sqrt{r})$ is at most

$$\frac{\log^2(m+n)+\frac{K\log^4(m+n)}{\sqrt{m+n}}}{2^k}\cdot(\sqrt{m+n}\frac{\sigma_s}{\delta_s})\left[(\mathcal{X}_0+\sigma)\cdot\max_{1\le i\le r}\|\mathbf{u}_{\pm i}\|_\infty+\frac{\|\mathcal{E}_0\|_{\text{op}}}{\sqrt{m+n}}\right],$$

where $\mathcal{X}_0\coloneqq\max_{1\le i\le r,1\le j\le m+n}|\mathbf{u}_{\pm i}^\top\mathcal{E}_0\mathbf{v}_j|$ with $\mathbf{v}_j, 1\le j\le m+n$ eigenvectors of $\mathcal{A}$. Define $X_0\coloneqq\max_{1\le i,j\le r}|u_i^\top E_0 v_j|$. As a result, we obtain that $\|\widetilde{A}_s-A_s\|_{2,\infty}$ is at most

$$\left[\log^2(m+n)+\frac{K\log^4(m+n)}{\sqrt{m+n}}\right]\cdot(\sqrt{m+n}\frac{\sigma_s}{\delta_s})\left[\mu\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right)(X_0+\sigma)+\frac{\|E_0\|_{\text{op}}}{\sqrt{m+n}}\right],$$

which improves the RHS of Theorem 4.2.

# D Proofs of the Technical Lemmas

## D.1 Bounding $G_{k,l}, l\ge 1$ (Proof of Lemma C.4)

We expand $\frac{1}{z^{l-1}}e_1^\top\mathcal{E}^l P$ into $\frac{1}{z^{l-1}}e_1^\top\mathcal{E}^l\left(\sum_{i=1}^r\frac{\sigma_i}{(z-\sigma_i)z}\mathbf{u}_{+i}\mathbf{u}_{+i}^\top+\sum_{i=1}^r\frac{-\sigma_i}{(z+\sigma_i)z}\mathbf{u}_{-i}\mathbf{u}_{-i}^\top\right)$, which can be rearranged as

$$\left[\sum_{i=1}^r\frac{\sigma_i(e_1^\top\mathcal{E}^l\mathbf{u}_{+i})}{z^l(z-\sigma_i)}\cdot\mathbf{u}_{+i}^\top\right]+\left[\sum_{i=1}^r\frac{-\sigma_i(e_1^\top\mathcal{E}^l\mathbf{u}_{-i})}{z^l(z+\sigma_i)}\cdot\mathbf{u}_{-i}^\top\right].$$

Since $\{\mathbf{u}_{\pm i}\}_{1\le i\le r}$ form an orthonormal system, $\|\frac{1}{z^{l-1}}e_1^\top \mathcal{E}^l P\|_2$ equals

$$\begin{aligned}&\tfrac{1}{|z|^l}\cdot\sqrt{\textstyle\sum_{i=1}^r \frac{\sigma_i^2\cdot(e_1^\top\mathcal{E}^l\mathbf{u}_{+i})^2}{|z-\sigma_i|^2}+\sum_{i=1}^r\frac{\sigma_i^2\cdot(e_1^\top\mathcal{E}^l\mathbf{u}_{-i})^2}{|z+\sigma_i|^2}}\\&\le\max_{1\le i\le r}|e_1^\top\mathcal{E}^l\mathbf{u}_{\pm i}|\cdot\frac{\sqrt{\sum_{i=1}^r\frac{\sigma_i^2}{|z-\sigma_i|^2}+\sum_{i=1}^r\frac{\sigma_i^2}{|z+\sigma_i|^2}}}{|z|^l}\end{aligned}\tag{D.1}$$

Hence,

$$\tfrac{1}{2\pi}\cdot\textstyle\int_\Gamma\left\|\frac{1}{z^{l-1}}e_1^\top\mathcal{E}^lP\right\|_2|\,\mathrm{d}z|\le\frac{\max_{1\le i\le r}|e_1^\top\mathcal{E}^l\mathbf{u}_{\pm i}|}{2\pi}\cdot\int_\Gamma\frac{\sqrt{\sum_{i=1}^r\frac{\sigma_i^2}{|z-\sigma_i|^2}+\sum_{i=1}^r\frac{\sigma_i^2}{|z+\sigma_i|^2}}}{|z|^l}|\,\mathrm{d}z|\,.\tag{D.2}$$

The second factor can be bounded as follows.

$$\begin{aligned}&\int_\Gamma\frac{\sqrt{\sum_{i=1}^r\frac{\sigma_i^2}{|z-\sigma_i|^2}+\sum_{i=1}^r\frac{\sigma_i^2}{|z+\sigma_i|^2}}}{|z|^l}|\,\mathrm{d}z|\le\sqrt{2r}\int_\Gamma\frac{1}{|z|^l}\cdot\max_{i\in[r]}\frac{\sigma_i}{|z\pm\sigma_i|}|\,\mathrm{d}z|\\&=\sqrt{2r}\int_{\Gamma^+}\frac{1}{|z|^l}\cdot\max_{i\in[r]}\frac{\sigma_i}{|z\pm\sigma_i|}|\,\mathrm{d}z|+\sqrt{2r}\int_{\Gamma^-}\frac{1}{|z|^l}\cdot\max_{i\in[r]}\frac{\sigma_i}{|z\pm\sigma_i|}|\,\mathrm{d}z|\,.\end{aligned}\tag{D.3}$$

Moreover,

$$\begin{aligned}\sqrt{2r}\int_{\Gamma^+}\frac{1}{|z|^l}\cdot\max_{i\in[r]}\frac{\sigma_i}{|z\pm\sigma_i|}|\,\mathrm{d}z|&\le\sqrt{2r}\frac{\sigma_s}{\delta_s/2}\cdot\int_{\Gamma^+}\frac{1}{|z|^l}|\,\mathrm{d}z|\\&\le\sqrt{2r}\frac{2\sigma_s}{\delta_s}\cdot\frac{1}{(\sigma_s/2)^{l-1}}\cdot3\log\left(\frac{10\sigma_1}{\sigma_s}\right).\end{aligned}\tag{D.4}$$

Here, the first inequality follows from the fact that $\frac{\sigma_i}{|z\pm\sigma_i|}\le\frac{\sigma_i}{|\mathcal{R}ez\pm\sigma_i|}\le\frac{\sigma_s}{\delta_s/2}$, while the last inequality is obtained by replacing $\delta_D$ by $\sigma_s/2$ in the bound in (Tran & Vu, 2026b, Appendix A). A similar upper bound holds for $\sqrt{2r}\int_{\Gamma^-}\frac{1}{|z|^l}\cdot\max_{i\in[r]}\frac{\sigma_i}{|z\pm\sigma_i|}|\,\mathrm{d}z|$.

Next, we handle the first factor as follows. We split $\mathcal{E}=\mathcal{E}_R+\mathcal{E}_0$, where $\mathcal{E}_R$ is a $(K,\sigma)$-bounded random matrix. By the triangle inequality, we have

$$\begin{aligned}|e_1^\top\mathcal{E}^l\mathbf{u}_{\pm i}|&=|e_1^\top(\mathcal{E}_R+\mathcal{E}_0)^l\mathbf{u}_{\pm i}|\\&\le|e_1^\top\mathcal{E}_R^l\mathbf{u}_{\pm i}|+\sum_{t=0}^{l-1}\binom{l}{t}\|\mathcal{E}_R\|_{\mathrm{op}}^t\|\mathcal{E}_0\|_{\mathrm{op}}^{l-t}\,.\end{aligned}\tag{D.5}$$

For the first term, by (Tran & Vu, 2025a, Lemma 4.2), we have

$$\begin{aligned}&\max_{1\le i\le r}|e_1^\top\mathcal{E}_R^l\mathbf{u}_{\pm i}|\le(2\sigma\sqrt{m+n})^l\\&\times\left(\left[\log(m+n)+\frac{K\log^3(m+n)}{\sqrt{m+n}}\right]\max_{1\le i\le r}\|\mathbf{u}_{\pm i}\|_\infty+\frac{\log^{3/2}(m+n)}{\sqrt{m+n}}\right),\end{aligned}\tag{D.6}$$

with probability at least $1-O((m+n)^{-2})$.

For the remaining terms, using Vu (2007); Bandeira & van Handel (2016), with probability at least $1-O((m+n)^{-2})$, we have

$$\binom{l}{t}\|\mathcal{E}_R\|_{\mathrm{op}}^t\|\mathcal{E}_0\|_{\mathrm{op}}^{l-t}\le\binom{l}{t}\cdot(2\sigma\sqrt{m+n})^t\|\mathcal{E}_0\|_{\mathrm{op}}^{l-t}\,.\tag{D.7}$$

Combining (D.2), (D.6), (D.7), (D.4), and the line below it, we obtain that $\frac{\int_\Gamma\left\|\frac{1}{z^{l-1}}e_1^\top\mathcal{E}^lP\right\|_2|\,\mathrm{d}z|}{2\pi C\sqrt{r}}$ is at most

$$\begin{aligned}&\left(\frac{4\sigma\sqrt{m+n}}{\sigma_s}\right)^{l-1}\cdot(\sigma\sqrt{m+n}\frac{\sigma_s}{\delta_s})\left[\log^2(m+n)+\frac{K\log^4(m+n)}{\sqrt{m+n}}\right]\max_{1\le i\le r}\|\mathbf{u}_{\pm i}\|_\infty\\&+\sum_{t=0}^{l-1}\binom{l}{t}\left(\frac{4\sigma\sqrt{m+n}}{\sigma_s}\right)^t\cdot\left(\frac{2\|\mathcal{E}_0\|_{\mathrm{op}}}{\sigma_s}\right)^{l-t-1}\cdot(\sqrt{m+n}\frac{\sigma_s}{\delta_s})\cdot\frac{\|\mathcal{E}_0\|_{\mathrm{op}}\log(m+n)}{\sqrt{m+n}}\\&\le\frac{1}{3^{l-1}}\cdot(\sigma\sqrt{m+n}\frac{\sigma_s}{\delta_s})\\&\times\left(\left[\log^2(m+n)+\frac{K\log^4(m+n)}{\sqrt{m+n}}\right]\max_{1\le i\le r}\|\mathbf{u}_{\pm i}\|_\infty+\frac{\|\mathcal{E}_0\|_{\mathrm{op}}\log(m+n)}{\sigma\sqrt{m+n}}\right),\end{aligned}$$

with probability at least $1 - O((m+n)^{-2})$ for some universal constant $C > 0$. The last inequality follows from the fact that $\sigma_s \geq \delta_s \geq 6(\|\mathcal{E}_R\|_{\mathrm{op}} + \|\mathcal{E}_0\|_{\mathrm{op}}) \geq 3\big[4\sigma\sqrt{m+n} + 2\|\mathcal{E}_0\|_{\mathrm{op}}\big]$.

This proves Lemma C.4.

### D.2 Bounding $G_{k,0}$ (Proof of Lemma C.5)

Consider the spectral decomposition of $\mathcal{A} := \sum_{i=1}^{n+m} \lambda_j \mathbf{v}_j \mathbf{v}_j^\top$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{m+n}$ are its eigenvalues with corresponding eigenvector $\mathbf{v}_j, 1 \leq j \leq m+n$. This gives another representation of $(zI - \mathcal{A})^{-1} = \sum_{j=1}^{m+n} \frac{\mathbf{v}_j \mathbf{v}_j^\top}{z - \lambda_j}$. Thus, we rewrite $z e_1^\top P \mathcal{E} (zI - \mathcal{A})^{-1}$ as follows.

$$\begin{aligned} z e_1^\top P \mathcal{E} (zI - \mathcal{A})^{-1} &= \left( \sum_{i=1}^{r} \frac{\sigma_i \mathbf{u}_{+i1}}{(z - \sigma_i)} \mathbf{u}_{+i}^\top + \sum_{i=1}^{r} \frac{-\sigma_i \mathbf{u}_{-i1}}{(z + \sigma_i)} \mathbf{u}_{-i}^\top \right) \mathcal{E} \left( \sum_{j=1}^{m+n} \frac{\mathbf{v}_j \mathbf{v}_j^\top}{z - \lambda_j} \right) \\ &= \sum_{j=1}^{m+n} \left( \sum_{i=1}^{r} \frac{\sigma_i \mathbf{u}_{+i1}}{(z - \sigma_i)} (\mathbf{u}_{+i}^\top \mathcal{E} \mathbf{v}_j) + \sum_{i=1}^{r} \frac{-\sigma_i \mathbf{u}_{-i1}}{(z + \sigma_i)} (\mathbf{u}_{-i}^\top \mathcal{E} \mathbf{v}_j) \right) \cdot \frac{\mathbf{v}_j^\top}{z - \lambda_j} \,. \end{aligned} \tag{D.8}$$

Therefore, $\|z e_1^\top P \mathcal{E} (zI - \mathcal{A})^{-1}\|_2$ equals

$$\sqrt{\sum_{j=1}^{m+n} \left| \sum_{i=1}^{r} \frac{\sigma_i \mathbf{u}_{+i1}}{(z - \sigma_i)} (\mathbf{u}_{+i}^\top \mathcal{E} \mathbf{v}_j) + \sum_{i=1}^{r} \frac{-\sigma_i \mathbf{u}_{-i1}}{(z + \sigma_i)} (\mathbf{u}_{-i}^\top \mathcal{E} \mathbf{v}_j) \right|^2 \cdot \frac{1}{|z - \lambda_j|^2}} \,,$$

which, by the Cauchy–Schwarz inequality, is at most

$$\sqrt{\sum_{j=1}^{m+n} 2r \left( \sum_{i=1}^{r} \left| \frac{\sigma_i \mathbf{u}_{+i1}}{(z - \sigma_i)} (\mathbf{u}_{+i}^\top \mathcal{E} \mathbf{v}_j) \right|^2 + \sum_{i=1}^{r} \left| \frac{-\sigma_i \mathbf{u}_{-i1}}{(z + \sigma_i)} (\mathbf{u}_{-i}^\top \mathcal{E} \mathbf{v}_j) \right|^2 \right) \cdot \frac{1}{|z - \lambda_j|^2}}$$

$$\leq \max_{1 \leq i \leq r, 1 \leq j \leq m+n} |\mathbf{u}_{\pm i}^\top \mathcal{E} \mathbf{v}_j| \cdot \max_{1 \leq i \leq r} \|\mathbf{u}_{\pm i}\|_\infty \sqrt{\sum_{j=1}^{m+n} 2r \left[ \sum_{i=1}^{r} \frac{\sigma_i^2}{|z - \sigma_i|^2} + \sum_{i=1}^{r} \frac{\sigma_i^2}{|z + \sigma_i|^2} \right] \frac{1}{|z - \lambda_j|^2}} \,.$$

Since $\frac{\sigma_i}{|z - \sigma_i|} \leq \frac{\sigma_i}{|\mathcal{R}ez - \sigma_i|} \leq \frac{\sigma_s}{\delta_s/2}$, $\|z e_1^\top P \mathcal{E} (zI - \mathcal{A})^{-1}\|_2$ is at most

$$\max_{1 \leq i \leq r, 1 \leq j \leq m+n} |\mathbf{u}_{\pm i}^\top \mathcal{E} \mathbf{v}_j| \cdot \max_{1 \leq i \leq r} \|\mathbf{u}_{\pm i}\|_\infty \cdot \frac{4r\sigma_s}{\delta_s} \frac{\sqrt{m+n}}{\min_j |z - \lambda_j|} \,.$$

Thus,

$$\begin{aligned} \int_\Gamma \left\| z e_1^\top P \mathcal{E} (zI - \mathcal{A})^{-1} \right\|_2 |\,\mathrm{d}z| \leq & \max_{\substack{1 \leq i \leq r \\ 1 \leq j \leq m+n}} |\mathbf{u}_{\pm i}^\top \mathcal{E} \mathbf{v}_j| \cdot \max_{1 \leq i \leq r} \|\mathbf{u}_{\pm i}\|_\infty \\ & \times \frac{4r\sigma_s \sqrt{m+n}}{\delta_s} \int_\Gamma \frac{1}{\min_j |z - \lambda_j|} |\,\mathrm{d}z| \,. \end{aligned}$$

By (Tran & Vu, 2026b, Appendix A), the RHS is at most

$$\max_{1 \leq i \leq r, 1 \leq j \leq m+n} |\mathbf{u}_{\pm i}^\top \mathcal{E} \mathbf{v}_j| \cdot \max_{1 \leq i \leq r} \|\mathbf{u}_{\pm i}\|_\infty \cdot \frac{4r\sigma_s \sqrt{m+n}}{\delta_s} \cdot 4 \log \left( \frac{10\sigma_1}{\delta_s} \right) \,.$$

On the other hand, by the Bernstein inequality, with probability at least $1 - O((m+n)^{-2})$,

$$\max_{1 \leq i \leq r, 1 \leq j \leq m+n} |\mathbf{u}_{\pm i}^\top \mathcal{E} \mathbf{v}_j| \leq \max_{1 \leq i \leq r, 1 \leq j \leq m+n} |\mathbf{u}_{\pm i}^\top \mathcal{E}_R \mathbf{v}_j| + |\mathbf{u}_{\pm i}^\top \mathcal{E}_0 \mathbf{v}_j| = O(\sigma \log(m+n) + \|\mathcal{E}_0\|_{\mathrm{op}}) \,.$$

Combining these estimates, with probability $1 - O((m+n)^{-2})$, there is a universal constant $C$ such that

$$\frac{\int_\Gamma \left\| z e_1^\top P \mathcal{E} (zI - \mathcal{A})^{-1} \right\|_2 |\,\mathrm{d}z|}{2\pi} \leq C\sqrt{r} \log^2(m+n)(\|\mathcal{E}_0\|_{\mathrm{op}} + \sigma) \cdot (\sqrt{m+n} \frac{\sigma_s}{\delta_s}) \cdot \max_{1 \leq i \leq r} \|\mathbf{u}_{\pm i}\|_\infty \,.$$

This proves our lemma.

### D.3 Bounding $\|E_R\|_{\mathrm{op}}$ (Proof of Lemma 4.1)

Recall that $E_R \coloneqq \widetilde{A}^{ub} - \mathbb{E}\,\widetilde{A}^{ub} + \widetilde{E}^{ub}$. Our goal is to bound $\|E_R\|_{\mathrm{op}}$. By the triangle inequality,

$$\|E_R\|_{\mathrm{op}} \le \|\widetilde{A}^{ub} - \mathbb{E}\,\widetilde{A}^{ub}\|_{\mathrm{op}} + \|\widetilde{E}^{ub}\|_{\mathrm{op}}\,. \tag{D.9}$$

Moreover, using Bandeira & van Handel (2016), we obtain

$$\|\widetilde{A}^{ub} - \mathbb{E}\,\widetilde{A}^{ub}\|_{\mathrm{op}} \le 6\cdot\left(\max\{M_R, M_C\} + \max_{i,j}\left|a_{ij}\sqrt{\tfrac{p_{ij}(1-p_{ij})}{p_i^2}}\right|\cdot\sqrt{\log(m+n)}\right)\,,\ \text{where}$$
$$M_R \coloneqq \max_{1\le i\le n}\sqrt{\textstyle\sum_{j=1}^m a_{ij}^2\,\frac{p_{ij}(1-p_{ij})}{p_i^2}}\ \text{and}\ M_C \coloneqq \max_{1\le j\le m}\sqrt{\textstyle\sum_{i=1}^n a_{ij}^2\,\frac{p_{ij}(1-p_{ij})}{p_i^2}}\,. \tag{D.10}$$

Since $K_A$ satisfies $\|A\|_\infty = \max_{i,j}|a_{ij}| \le K_A$, we have

$$M_R \le K_A\cdot\max_{1\le i\le n}\sqrt{\textstyle\sum_{j=1}^m \frac{p_{ij}(1-p_{ij})}{p_i^2}} \le K_A\cdot\max_{1\le i\le n}\sqrt{\textstyle\sum_{j=1}^m \frac{p_{ij}}{p_i^2}} = K_A\cdot\sqrt{\tfrac{m}{\min_{1\le i\le n} p_i}}\,.$$

Similarly,

$$M_C \le K_A\cdot\max_{1\le j\le m}\sqrt{\textstyle\sum_{i=1}^n \frac{p_{ij}(1-p_{ij})}{p_i^2}} \le K_A\cdot\sqrt{\frac{\max_{1\le j\le m}\sum_{i=1}^n \frac{p_{ij}}{p_i}}{\min_{1\le i\le n} p_i}}\,.$$

Therefore,

$$\max\{M_R, M_C\} \le K_A\cdot\sqrt{\tfrac{m+r_p n}{\min p_i}}\,.$$

Since

$$\max_{i,j}\left|a_{ij}\sqrt{\tfrac{p_{ij}(1-p_{ij})}{p_i^2}}\right| \le K_A\cdot\max_{i,j}\sqrt{\tfrac{p_{ij}/p_i}{p_i}} \le K_A\cdot\sqrt{\tfrac{r_p}{\min p_i}}\,,$$

we obtain

$$\|\widetilde{A}^{ub} - \mathbb{E}\,\widetilde{A}^{ub}\|_{\mathrm{op}} \le 12K_A\sqrt{\log(m+n)}\cdot\sqrt{\tfrac{m+r_p n}{\min p_i}}\,. \tag{D.11}$$

By a similar argument, we also obtain that $\frac{\|\widetilde{E}^{ub}\|_{\mathrm{op}}}{6}$ is at most

$$\max_{\substack{1\le i\le n\\ 1\le j\le m}}\left\{\sqrt{\textstyle\sum_{k=1}^m \mathbb{E}\,\varepsilon_{ik}^2\,\frac{p_{ik}}{p_i^2}}, \sqrt{\textstyle\sum_{l=1}^n \mathbb{E}\,\varepsilon_{lj}^2\,\frac{p_{lj}}{p_l^2}}\right\} + \max_{i,j}\left|K_E\sqrt{\tfrac{p_{ij}}{p_i^2}}\right|\sqrt{\log(m+n)}\,,$$

and hence

$$\|\widetilde{E}^{ub}\|_{\mathrm{op}} \le 12K_E\sqrt{\log(m+n)}\cdot\sqrt{\tfrac{m+r_p n}{\min p_i}}\,. \tag{D.12}$$

Combining the inequalities (D.11) and (D.12) with (D.9), we finally obtain

$$\|E_R\|_{\mathrm{op}} \le 12(K_A + K_E)\left(\sqrt{\log(m+n)}\cdot\sqrt{\tfrac{m+r_p n}{\min p_i}}\right)\,. \tag{D.13}$$

This completes the proof of Lemma 4.1.