Linear Regression with Self-Selection Bias in Polynomial Time

Felix Zhou
Yale University
felix.zhou@yale.edu

Abstract

We revisit the problem of estimating k linear regressors with self-selection bias in n dimensions up to error ε with the maximum self-selection criterion, as introduced by Cherapanamjeri, Daskalakis, Ilyas, and Zampetakis [CDIZ23, STOC]. Our main result is a $\mathrm{poly}(n, k, 1/\varepsilon)$ time algorithm for the unknown-index variant of this problem, thereby improving upon the running times of the best known algorithms due to Gaitonde and Mossel [GM24] and Kalavasis, Mehrotra, and Zhou [KMZ26, COLT] with exponential dependence on k.

We depart from the moment testing paradigm of prior work, which exhaustively searches over a k-dimensional subspace spanned by the regressors. Instead, our algorithm is driven by an improved understanding of the moments of the self-selection problem. We show that, restricting to observations with large response values, the degree-$(\log n)$ Hermite moment of the covariates approximately decomposes as a weighted sum of orthogonal rank-one tensors of the regressors. The first step of our algorithm can be viewed as a tensor decomposition algorithm for these implicitly represented high-order tensors, which outputs unit vectors parallel to the regressors. A secondary step based on a simple variant of ordinary least squares extracts estimates for the regressors given these unit vectors. We believe our analytic insight and algorithmic components may have applications to designing efficient algorithms for other latent-variable problems.

Contents

1 Introduction **1**
- 1.1 Our Contributions . . . 2
  - 1.1.1 Technical Contributions . . . 3

2 Technical Overview **3**
- 2.1 Prior Techniques . . . 3
- 2.2 Our Approach . . . 4
  - 2.2.1 Hermite Moments of Covariates . . . 4
  - 2.2.2 Projections of Hermite Tensors . . . 6
  - 2.2.3 Observable-Operator Learning . . . 8
  - 2.2.4 Restricted Least-Squares . . . 10
  - 2.2.5 Moment Peeling . . . 11
- 2.3 Organization . . . 12

3 Related Works **13**

4 Preliminaries **15**
- 4.1 Hyperparameters . . . 15
- 4.2 Notation . . . 15
- 4.3 Properties of Gaussians . . . 16
- 4.4 Hermite Polynomials . . . 16
- 4.5 Spectral Perturbation Bounds . . . 17
- 4.6 Heavy-Tail Mean Estimation . . . 17

5 Structure of Restricted Hermite Moments **18**
- 5.1 Decomposition of the Restricted Hermite Moments . . . 20
  - 5.1.1 Single-Response Moments . . . 20
  - 5.1.2 Moment Decomposition Error . . . 20
- 5.2 Bounds on the Component Weights . . . 22
  - 5.2.1 Restricted Moment Coefficients . . . 22
  - 5.2.2 Slice Probabilities and Weights . . . 23
  - 5.2.3 Weights Across Degrees . . . 25
- 5.3 Proof of Theorem 5.1 . . . 25

6 Implicit Tensor Decomposition **28**
- 6.1 Proof Roadmap . . . 29
- 6.2 Iterative Projections . . . 32
  - 6.2.1 Representation of Projection Maps . . . 32
  - 6.2.2 Mean Estimation Oracle . . . 33
  - 6.2.3 Cumulative Projection Error . . . 33
  - 6.2.4 Proof of Theorem 6.3 . . . 37
- 6.3 Observable-Operator Learning . . . 38
  - 6.3.1 Proof of Theorem 6.5 . . . 38

6.3.2 Proof of Theorem 6.6 . . . 43
6.4 Proof of Theorem 6.1 . . . 46

7 Restricted Least-Squares **48**
7.1 Proof of Theorem 7.1 . . . 49

8 Moment Peeling & Estimation **54**
8.1 Proof Roadmap . . . 55
8.1.1 Signed Peeling Weight . . . 56
8.1.2 Projected Hermite Features . . . 56
8.1.3 Error Decomposition . . . 57
8.2 Sampling Error . . . 59
8.3 Recovered-Regressor Error . . . 61
8.4 Peeling Bias . . . 62
8.5 Proof of Theorem 8.1 . . . 63

9 Proof of Theorem 1.3 **66**

A Additional Preliminaries **75**
A.1 Gaussian Moments and Elementary Bounds . . . 75
A.2 Linear Algebra . . . 76

B Deferred Proofs for Section 4 **76**
B.1 Hermite Polynomials . . . 76
B.2 Heavy-Tail Mean Estimation . . . 77

C Deferred Proofs for Section 5 **78**
C.1 Weights and Scaled Regressors . . . 78
C.2 Single-Response Moments . . . 78
C.3 Moment Decomposition Error . . . 79

D Deferred Proofs for Section 6 **81**
D.1 Iterative Projections . . . 81
D.2 Observable-Operator Learning . . . 82
D.3 Proof of Theorem 6.1 . . . 87

E Deferred Proofs for Section 7 **88**

F Deferred Proofs for Section 8 **89**
F.1 Second-Moment Bounds . . . 90
F.2 Sampling Error . . . 91
F.3 Peeling Bias . . . 91

G Deferred Proofs for Section 9 **94**

H AI Methodology **95**

1 Introduction

In their seminal work on disequilibrium econometrics, Fair and Jaffee [FJ72] studied a linear model of the housing market where there is a supply $S(\boldsymbol{x}, \xi_S) = \boldsymbol{w}_S^\top \boldsymbol{x} + \xi_S$ and a demand $D(\boldsymbol{x}, \xi_D) = \boldsymbol{w}_D^\top \boldsymbol{x} + \xi_D$ of houses with feature $\boldsymbol{x}$, but supply does not equal demand. The actual quantity transacted is $Q(\boldsymbol{x}) = \min\{S(\boldsymbol{x}, \xi_S), D(\boldsymbol{x}, \xi_D)\}$ where (ξ_S, ξ_D) are random noise. A statistician aiming to estimate the regressors $\boldsymbol{w}_D, \boldsymbol{w}_S$ from observations $(\boldsymbol{x}, Q(\boldsymbol{x}))$ faces difficulties with vanilla linear regression methods:

(1) we do not observe whether a given transaction is constrained by supply or demand, and

(2) even if this is known, partitioning the data and running an ordinary least squares (OLS) estimator on all the supply-constrained data to estimate $\boldsymbol{w}_S$ and a separate OLS estimator on all the demand-constrained data to estimate $\boldsymbol{w}_D$ will likely yield biased estimates.

To see the second point, note that the statistician will never observe the demand $D(\boldsymbol{x}, \xi_D)$ for houses with low supply $S(\boldsymbol{x}, \xi_S) \ll D(\boldsymbol{x}, \xi_D)$ and vice versa. This leads to biased observations in the partitioned datasets which in turn induce biases in the OLS estimators on those datasets.

Unknown-Index Maximum Self-Selection. The example above is an instance of linear regression with *self-selection* bias. In the general setting, we observe features $\boldsymbol{X}$ accompanied by an outcome $Y \in \{Y_1, \ldots, Y_k\}$ selected from the responses of k underlying linear models: $Y_i = \boldsymbol{w}_i^\top \boldsymbol{X} + \xi_i$. In this work, we focus on the maximum self-selection mechanism, for which the observed response is $Y_{\max} := \max_{i \in [k]} Y_i$.[1] Importantly, to capture the setting of Fair and Jaffee [FJ72], we do *not* assume the index of the selected model is known. This is also known as the *unknown-index* setting.

More broadly, self-selection also arises in occupational choice, where workers choose among occupations with different potential earnings and only the earnings from the chosen occupation are observed [Roy51]. In observational medicine, treatment choices depend on prognostic information, which can bias comparisons of outcomes [GN80; RR83]. A related inverse problem arises in inverse reinforcement learning, where one seeks a reward function from an expert's observed behavior [NR00]. Other examples arise in auctions and voluntary disclosure. Auction data may reveal only the transaction price, leaving the winning bidder and losing bids unobserved [AH02]. Finally, models of voluntary disclosure study how informed sellers choose what information about product quality to reveal [Gro81; Mil81].

Computational and Statistical Efficiency. Despite the long history and use of self-selection models, algorithms with provable guarantees for finite samples have only recently been studied. Cherapanamjeri, Daskalakis, Ilyas, and Zampetakis [CDIZ23] initiated the study of sample- and time-efficient algorithms for linear regression under the maximum self-selection bias with k regressors in the unknown-index setting. If n is the ambient dimension, they gave an algorithm with $\mathrm{poly}(n)\exp(\mathrm{poly}(k))$ sample complexity and running time that achieves $\mathrm{poly}(1/k)$ error.

[1] The minimum in the model of Fair and Jaffee [FJ72] becomes a maximum after negating the observed response, parameter vectors, and noises.

The algorithm of Gaitonde and Mossel [GM24] uses $\mathrm{poly}(n, k, 1/\varepsilon)$ samples and runs in time $\mathrm{poly}(n, k) + (1/\varepsilon)^{\widetilde{O}(k)}$. Using this algorithm to supply a warm start to the local-convergence algorithm of Kalavasis, Mehrotra, and Zhou [KMZ26] further improves the running time to $\mathrm{poly}(n, k, 1/\varepsilon) + 2^{\widetilde{O}(k)}$.

Henceforth, **we refer to unknown-index linear regression with maximum self-selection mechanism as the *self-selection problem***. Our main result is the first algorithm for the self-selection problem with both $\mathrm{poly}(n, k, 1/\varepsilon)$ sample complexity and running time, improving on the time complexity of all prior works.

1.1 Our Contributions

Before stating our results, let us introduce the formal model of the self-selection problem.

Definition 1.1 (Maximum Self-Selection; [CDIZ23])**.** Fix integers $n \geq 1$ and $2 \leq k \leq n$.[2] We wish to recover unknown regressors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k \in \mathbb{R}^n$ and we observe independent samples $(\boldsymbol{X}, Y_{\max})$, where

$$\begin{aligned} \boldsymbol{X} &\sim N(\boldsymbol{0}, I_n)\,, \qquad Y_{\max} = \max_{i \in [k]} Y_i\,, \\ Y_i &= \boldsymbol{w}_i^\top \boldsymbol{X} + \xi_i\,, \qquad \xi_i \sim_{i.i.d.} N(0,1)\,. \end{aligned} \tag{1}$$

The noise vector $\boldsymbol{\xi} = (\xi_i)_{i \in [k]}$ is independent of $\boldsymbol{X}$. Note that the maximizing index is *not* observed.

In order to design algorithms with finite-sample guarantees, we make some standard geometric assumptions introduced in [CDIZ23] and shown to be necessary in [GM24]. Let $\|\cdot\|_2$ denote the Euclidean norm.

Assumption 1.2 (Boundedness and Uncoveredness)**.** For some absolute constants $B \geq 1$ and $\Delta \in (0, 1]$, the regressors satisfy

$$\Delta \leq \|\boldsymbol{w}_i\|_2 \leq B, \qquad \left|\langle \boldsymbol{w}_i, \boldsymbol{w}_j \rangle\right| \leq \min\{\|\boldsymbol{w}_i\|_2^2, \|\boldsymbol{w}_j\|_2^2\} - \Delta^2 \quad (i \neq j). \tag{2}$$

A simple consequence of Assumption 1.2 that we use repeatedly is the angular separation bound

$$\frac{\left|\langle \boldsymbol{w}_i, \boldsymbol{w}_j \rangle\right|}{\|\boldsymbol{w}_i\|_2 \, \|\boldsymbol{w}_j\|_2} \leq 1 - \frac{\Delta^2}{B^2} \qquad (i \neq j). \tag{3}$$

We are now ready to formally state our main recovery result. As the self-selection model is invariant under permutation, we recover the regressors up to some permutation.

Theorem 1.3 (Polynomial-Time Self-Selection)**.** *There is a randomized algorithm with the following guarantee. Given $k \geq 2$, desired error $\varepsilon \in (0, 1)$, failure probability $\delta \in (0, 1/3)$, and sample access to an instance of the self-selection model (Definition 1.1) satisfying boundedness and uncoveredness conditions (Assumption 1.2) with constants Δ, B, the algorithm returns $\widehat{\boldsymbol{w}}_1, \ldots, \widehat{\boldsymbol{w}}_k \in \mathbb{R}^n$ such that, with probability*

[2] The case $k = 1$ is just ordinary linear regression. By padding the covariate if necessary, we may assume without loss of generality that $k \leq n$.

at least $1-\delta$*, there is a permutation* π *of the indices under which* $\max_i \left\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_{\pi(i)}\right\|_2 \le \varepsilon$*. Its sample complexity and running time are both*

$$\widetilde{O}\left(\left(\frac{nk}{\varepsilon}\right)^{\mathrm{poly}(B/\Delta)} \log(1/\delta)\right).$$

As we mentioned, this yields the first polynomial-time algorithm for the self-selection problem.

1.1.1 Technical Contributions

Our algorithmic improvements are driven by an improved understanding of the moments of the self-selection problem. The key insight is that, conditional on large responses, the conditional distribution of the *covariate* is essentially a k-mixture distribution over the k regressors (see Figure 1). Mathematically, the higher-order conditional Hermite moments approximately decompose as a weighted sum of nearly orthogonal rank-one tensors. This departs from the conditional moment-testing paradigm from prior work, which exhaustively searches over a k-dimensional subspace spanned by the k regressors using conditional moments of the *response* $Y_{\max}$.

To exploit this structure algorithmically, we must handle uneven weights and biases that occur in the decomposition above. By carefully analyzing novel combinations of tensor and spectral methods from the literature on mixture models [AHK12; HK13; LL24; DK25], we are able to extract unit vectors $\widehat{\boldsymbol{u}}_i \approx \boldsymbol{w}_i / \|\boldsymbol{w}_i\|_2$ parallel to the regressors $\boldsymbol{w}_i$'s. Post-processing by a simple variant of ordinary least squares then yields accurate estimates of the regressors.

2 Technical Overview

We now present the key technical ingredients of our algorithm. For the sake of clarity, it is easiest to explain our algorithm in terms of true population expectations and ignore both sampling error and small bias terms. We will also treat K as an absolute constant that disappears in the big-O notation.

Notation. In this overview, we use bold letters for vectors, such as $\boldsymbol{X}, \boldsymbol{w}_i, \boldsymbol{b}$; ordinary capital letters for matrices, such as $A, C_r, M(\boldsymbol{b})$; and sans-serif capital letters for tensors, such as $\mathsf{H}_d(\boldsymbol{X})$ and R_d. Write $\boldsymbol{\xi} = (\xi_i)_{i\in[k]}$ and $\boldsymbol{\xi}_I = (\xi_i)_{i\in I}$ for $I \subseteq [k]$. For a positive integer m, write $[m] = \{1, \dots, m\}$. We use $\langle\cdot,\cdot\rangle$ for the Euclidean inner product, and $\|\cdot\|_2, \|\cdot\|_F, \|\cdot\|_{\mathrm{op}}$ for the Euclidean, Frobenius, and operator norms. The symbol $\mathbf{1}\{\mathcal{E}\}$ denotes the indicator of an event $\mathcal{E}$.

We use $\otimes$ for tensor products and $\boldsymbol{v}^{\otimes d}$ for the d-fold tensor product of $\boldsymbol{v}$. The space $\mathrm{Sym}^d(\mathbb{R}^n)$ consists of symmetric order-d tensors on $\mathbb{R}^n$. The notation $\mathsf{A} \lrcorner \boldsymbol{b}$ denotes contraction against $\boldsymbol{b}$ in the last tensor factor. We write $\mathrm{i} = \sqrt{-1}$ and use Re for the entrywise real part. Finally, Φ denotes the standard Gaussian CDF and ϕ_{σ^2} the density of $N(0, \sigma^2)$.

2.1 Prior Techniques

Moments of Responses. Before we begin, it is instructive to recall prior approaches.

Cherapanamjeri, Daskalakis, Ilyas, and Zampetakis [CDIZ23] test candidate directions using higher-order moments of the response $Y_{\max}$. They restrict to samples whose projected covariates lie in a thin tube around the candidate line. This gives rise to an exhaustive search algorithm that tests every member of an ε-net using conditional moments. A straightforward implementation would require $(1/\varepsilon)^{\Omega(n)}$ samples and running time, but the authors were able to reduce this to $\mathrm{poly}(n)\exp(\mathrm{poly}(k))$ samples and running time by restricting the search space to a k-dimensional subspace approximately spanned by the regressors.

Building on these ideas, Gaitonde and Mossel [GM24] test a candidate $\widehat{w}$ by conditioning on a large projection of $\boldsymbol{X}$ along the candidate direction. This event has inverse-polynomial probability, while the tube in the earlier test can have exponentially small probability in k. They only need conditional moments up to degree two. As a consequence, empirical estimates of these moments using samples concentrate and they can reuse samples across tests. This reduces the sample complexity to $\mathrm{poly}(n, k, 1/\varepsilon)$, but the running time remains $\mathrm{poly}(n, k, 1/\varepsilon) + (1/\varepsilon)^{\widetilde{O}(k)}$.

Local Convergence. Kalavasis, Mehrotra, and Zhou [KMZ26] showed that stochastic gradient descent on the natural log-likelihood function of observations converges, assuming the initialization point is within $\mathrm{poly}(1/k)$-distance of the true regressors. Combined with the previous algorithms for the warm start, their algorithm maintains the $\mathrm{poly}(n, k, 1/\varepsilon)$ sample complexity and improves the running time to $\mathrm{poly}(n, k, 1/\varepsilon) + 2^{\widetilde{O}(k)}$.

In this work, we design an algorithm with $\mathrm{poly}(n, k, 1/\varepsilon)$ sample complexity and running time, improving on the running time of all prior works. As part of our algorithm we will design a local convergence subroutine that only requires the initialization point to be within $O(1)$-distance of the true regressors, improving on the convergence radius of [KMZ26].

2.2 Our Approach

We now present an overview of our techniques. As mentioned, it is easiest to state our algorithm in terms of exact expectations and ignore small bias terms. The actual analysis uses samples to estimate this expectation and carefully track the bias. We defer these details to the respective technical sections.

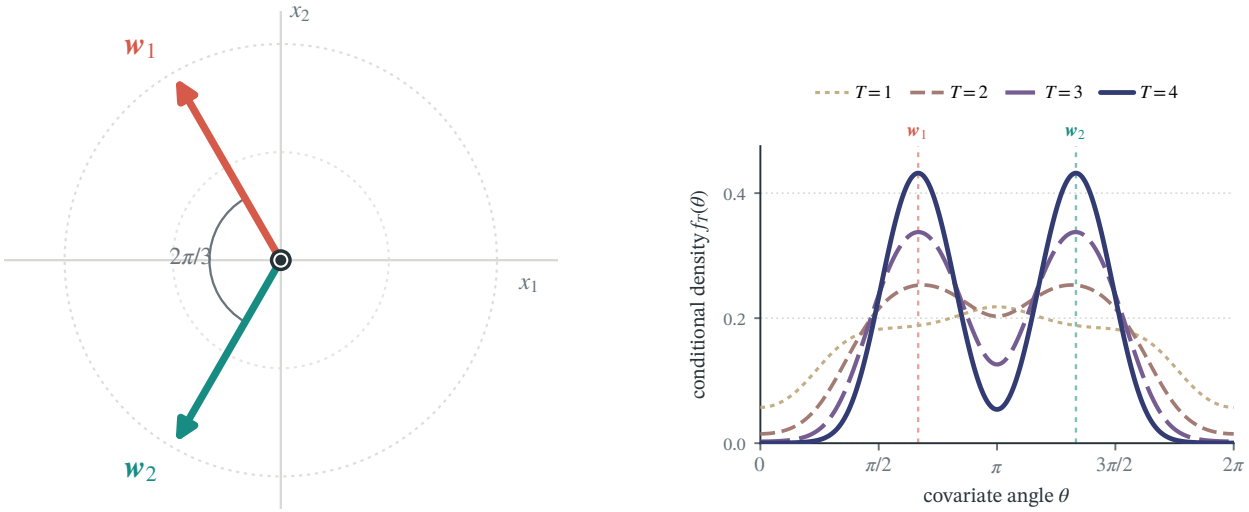
We first show that restricting the maximum response to a tail slice exposes the regressor directions in the Hermite moments of the covariates (Section 2.2.1). We then explain how to project these moments without forming the full tensors (Section 2.2.2) and use observable-operator learning to recover a block of approximate regressor directions (Section 2.2.3). Restricted least squares refines these directions into estimates of the full regressors (Section 2.2.4). Finally, moment peeling lets us estimate the moments of the remaining regressors from the original observations and repeat the procedure (Section 2.2.5).

2.2.1 Hermite Moments of Covariates

The focus of prior works is to study the conditional distribution of the *response* $Y_{\max}$. In particular, conditioned on the covariate $\boldsymbol{w}_i^\top \boldsymbol{X} = t \geq \Omega(\sqrt{\log n})$ being highly correlated with a regressor $\boldsymbol{w}_i$,

we expect the maximum response to be attained by Y_i, $(Y_{\max} \mid \boldsymbol{w}_i^\top \boldsymbol{X} = t) = (Y_i \mid \boldsymbol{w}_i^\top \boldsymbol{X} = t)$, except with $\mathrm{poly}(1/n)$ probability.

Our key observation is that the converse intuition also holds. That is, conditioned on the response $Y_{\max} = T \geq \Omega(\sqrt{\log n})$ being large, we expect the *covariates* to be highly correlated with some regressor. In particular, the conditional distribution of the covariates $\boldsymbol{X} \mid Y_{\max} = T$ should be a k modal distribution about the regressors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_k$ (see Figure 1). We formalize this intuition through the restricted Hermite moments in Section 5 and highlight the key ideas below.



(a) Regressors

(b) Conditional angle density

Figure 1: For two regressors $\boldsymbol{w}_1, \boldsymbol{w}_2 \in \mathbb{R}^2$ separated by angle $2\pi/3$ (a), the density of the covariate angle conditioned on $Y_{\max} \in [T, T+1]$ concentrates around the regressor directions as T grows (b).

Ordinary Linear Regression ($k = 1$). To make this more precise, let us study the conditional moments of $\boldsymbol{X} \mid Y_{\max} = T$. More specifically, we study the degree-d conditional *Hermite tensors* $\mathsf{H}_d(\boldsymbol{X}) \in \mathrm{Sym}^d(\mathbb{R}^n)$ of the covariate $\boldsymbol{X} \sim N(\boldsymbol{0}, I_n)$ (see Section 4.4). Recall that the (univariate) *Hermite polynomials* $\mathrm{He}_d(x) : \mathbb{R} \to \mathbb{R}, d \geq 0$ are the unique family of monic polynomials that are orthogonal under Gaussian measure. Their higher-dimensional analogues are the Hermite tensors $\mathsf{H}_d(\cdot) : \mathbb{R}^n \to \mathrm{Sym}^d(\mathbb{R}^n)$, $d \geq 0$, taking values in a space of dimension $\dim \mathrm{Sym}^d(\mathbb{R}^n) = \binom{n+d-1}{d}$.

We first consider ordinary linear regression (OLS) as a special case of $k = 1$ with $\boldsymbol{X} \sim N(\boldsymbol{0}, I_n)$, $Y = \boldsymbol{w}^\top \boldsymbol{X} + \xi$ with $\xi \sim N(0,1)$. Let $\phi_{\sigma^2}(t)$ denote the density of $N(0, \sigma^2)$ and $\boldsymbol{u} := \boldsymbol{w}/\|\boldsymbol{w}\|_2$ denote the unit vector parallel to the regressor. From a classic identity for the restricted Hermite tensors of the OLS covariate [Lan57; WN14; Sev26],

$$\mathbb{E}_{\boldsymbol{X},\xi}[\mathsf{H}_d(\boldsymbol{X})\mathbf{1}\{T \leq Y \leq T+1\}] = \lambda^{(d)}(\|\boldsymbol{w}\|_2)\boldsymbol{u}^{\otimes d}, \qquad \text{where} \quad \lambda^{(d)} = \lambda^{(d)}(\|\boldsymbol{w}\|_2) > 0\,.$$

Thus, up to a coefficient $\lambda^{(d)} = \lambda^{(d)}(\|\boldsymbol{w}\|_2)$ that depends on the norm of $\boldsymbol{w}$, the d-th restricted moment of the covariate is precisely the degree-d tensor power of $\boldsymbol{u}$. This identity can be proved using the orthogonal decomposition of $\boldsymbol{X}$ along $\boldsymbol{u}$ and the fact that every positive-degree Hermite polynomial has mean zero under a standard Gaussian.

Regression with Self-Selection ($k \geq 2$). Let us now consider the general case of $k \geq 2$ and replace Y with the maximum $Y_{\max}$. Let $I = I(Y_1, \ldots, Y_k) \in [k]$ denote the index of the maximizer response. By decomposing the indicator, we see that

$$\mathbf{1}\{T \leq Y_{\max} \leq T+1\} = \sum_{i \in [k]} \mathbf{1}\{I = i\}\mathbf{1}\{T \leq Y_i \leq T+1\} = \sum_{i \in [k]} (1 - \mathbf{1}\{I \neq i\})\mathbf{1}\{T \leq Y_i \leq T+1\}\,.$$

Write $\lambda_i^{(d)} = \lambda^{(d)}(\|\boldsymbol{w}_i\|_2)$ for the degree-d weight of the i-th regressor. Using this decomposition and the special case of $k = 1$, we see that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X},\xi}[\mathsf{H}_d(\boldsymbol{X})\mathbf{1}\{T \leq Y_{\max} \leq T+1\}] \\
&\quad = \sum_{i \in [k]} \lambda_i^{(d)} \boldsymbol{u}_i^{\otimes d} - \underbrace{\sum_{i \in [k]} \mathbb{E}_{\boldsymbol{X},\xi}[\mathsf{H}_d(\boldsymbol{X})\mathbf{1}\{I \neq i\}\mathbf{1}\{T \leq Y_i \leq T+1\}]}_{=:\mathsf{R}_d}\,.
\end{aligned}$$

In other words, up to a remainder term R_d, the degree-d restricted Hermite moment is precisely a weighted sum of degree-d tensor powers of the normalized regressors! Let us analyze some properties of this decomposition.

Near Orthogonality. A key property of this decomposition is that the rank-one tensors are nearly orthogonal. Indeed, by the separation conditions (Assumption 1.2),

$$\left|\left\langle \boldsymbol{u}_i^{\otimes s}, \boldsymbol{u}_j^{\otimes s} \right\rangle\right| = \left|\left\langle \frac{\boldsymbol{w}_i}{\|\boldsymbol{w}_i\|_2}, \frac{\boldsymbol{w}_j}{\|\boldsymbol{w}_j\|_2} \right\rangle\right|^s \leq \left(1 - \frac{\Delta^2}{B^2}\right)^s = e^{-\Omega(s)}\,.$$

Choosing $s = \Theta(\log n)$ makes this inner product at most $\mathrm{poly}(1/n)$.

Small Bias. Another important feature is that the bias term R_d is small. Fix an index $i \in [k]$. The remainder event $\mathbf{1}\{I \neq i\}\mathbf{1}\{T \leq Y_i \leq T+1\}$ occurs when $Y_i \geq T = \Theta(\sqrt{\log n})$ but does not attain the maximum. Intuitively, if every other Y_j was independent of Y_i, then by a union bound, the remainder event occurs with probability at most $\sum_{j \neq i} \Pr[Y_j > T \mid Y_i = T] \leq ke^{-\Omega(T^2)} = \mathrm{poly}(1/n)$. While this is too good to be true, we can apply the separation conditions to show that $\mathbb{E}[Y_j \mid Y_i = T] \leq cT$ for some absolute constant $c \in (0,1)$ and that the conditional variance remains constant. Hence we can still prove a similar tail bound on the remainder event. Combined with a careful bound on the variance of the (unrestricted) degree-s Hermite moment, this shows that the remainder $\|\mathsf{R}_d\|_F = \mathrm{poly}(1/n)$ is indeed negligible.

For the rest of the overview, we drop the remainder terms and treat the nearly orthogonal lifted vectors as orthonormal, with the understanding that both errors are accounted for in the formal proofs. Next, we will see how to exploit the structural properties above to design efficient algorithms for self-selection.

2.2.2 Projections of Hermite Tensors

In order to exploit the structure of the high-order tensors to design a polynomial time algorithm, we must avoid materializing the tensors explicitly. Perhaps a natural attempt is to project the tensors into a lower-dimensional space and perform the necessary computation there. In Section 6.2, we analyze an iterative projection method that avoids explicitly forming the higher-order tensors.

Implicitly Accessing Higher-Order Tensors. Our algorithm builds on the works of Li and Liu [LL24] and Diakonikolas and Kane [DK25] that use iterative projections to compute projections of high-order tensors. This enables them to evaluate inner products of these tensors without explicitly computing their entries. We were unable to directly apply their framework for a few technical reasons. One issue is that our restricted Hermite tensors do not cleanly decompose as a sum of rank-one tensors and have some irreducible error R_d. Another issue is that the coefficients $\lambda_i^{(d)}$ depend on the degree d, whereas the framework of Diakonikolas and Kane [DK25] assumes a sequence of moment tensors with the same weights across degrees. Most importantly, we were unable to reduce the problem of recovering the regressors to evaluating inner products. Nonetheless, we can borrow the idea of iterative projections for dimensionality reduction in our setting, although we need to go through a new analysis that handles the irreducible bias R_d.

Complex Hermite Outer Products. The iterative projection algorithm is easiest to state if we think of our Hermite tensors as flattened matrices. For this purpose, it is convenient to introduce an alternative representation of the Hermite tensors by lifting the covariates to $\mathbb{C}^n$. Indeed, for each covariate $\boldsymbol{X}$, independently draw an auxiliary Gaussian vector $\boldsymbol{X}' \sim N(\mathbf{0}, I_n)$ and form the complexified covariate $\widetilde{\boldsymbol{X}} := \boldsymbol{X} + \mathrm{i}\boldsymbol{X}'$. By the Gaussian integral representation of Hermite polynomials (see Section 4.4), we have $\mathsf{H}_d(\boldsymbol{X}) = \frac{1}{\sqrt{d!}}\mathbb{E}_{\boldsymbol{X}'}[\widetilde{\boldsymbol{X}}^{\otimes d} \mid \boldsymbol{X}]$. Let $\mathrm{Mat}_r : (\mathbb{R}^n)^{\otimes 2d} \to \mathbb{R}^{n^d} \times \mathbb{R}^{n^d}$ denote the flattening maps. This gives the identity

$$\mathrm{Mat}_r\big(\mathsf{H}_{2r}(\boldsymbol{X})\big) = \mathbb{E}_{\boldsymbol{X}'}\left[\frac{\mathrm{Re}[\widetilde{\boldsymbol{X}}^{\otimes r}(\widetilde{\boldsymbol{X}}^{\otimes r})^\top]}{\sqrt{(2r)!}} \;\middle|\; \boldsymbol{X}\right],$$

where $\mathrm{Re}[\cdot]$ denotes the real part of the complex argument. Note the identity holds even when fixing the real part $\boldsymbol{X}$ so that it still holds under restrictions of $\boldsymbol{X}$.

Thus for any linear projection $C : \mathbb{R}^{n^r} \to \mathbb{R}^p$, we can compute projections of the flattened restricted Hermite tensors as

$$C\,\mathbb{E}_{\boldsymbol{X},\xi}\left[\mathrm{Mat}_r(\mathsf{H}_{2r}(\boldsymbol{X}))\mathbf{1}\{T \le Y_{\max} \le T+1\}\right]C^\top = \mathbb{E}_{\boldsymbol{X},\xi,\boldsymbol{X}'}\left[\frac{\mathrm{Re}[C\widetilde{\boldsymbol{X}}^{\otimes r}(C\widetilde{\boldsymbol{X}}^{\otimes r})^\top]}{\sqrt{(2r)!}}\mathbf{1}\{T \le Y_{\max} \le T+1\}\right] \tag{4}$$

$$\approx \sum_{i\in[k]} \lambda_i^{(2r)} \left[(C\boldsymbol{u}_i^{\otimes r})\right]\left[(C\boldsymbol{u}_i^{\otimes r})\right]^\top. \tag{5}$$

In other words, if we can compute $C\widetilde{\boldsymbol{X}}^{\otimes r}$ in polynomial time, we can compute the lower-dimensional projections of the flattened matrices without explicitly forming them in high dimensions beforehand!

Iterative Projections. We now present the iterative projection method due to Li and Liu [LL24] that implicitly computes projections of flattened Hermite tensors.

Our goal is to implicitly compute $C_s\widetilde{\boldsymbol{X}}^{\otimes s}$ such that $\mathrm{span}\{\boldsymbol{u}_i^{\otimes s} : i \in [k]\} \subseteq \mathrm{range}(C_s^\top)$. We begin with the trivial base case of the degree $r = 1$. We can simply take the degree-1 projection matrix to be $C_1 = I_n$. Now, assume that we have computed $C_r\widetilde{\boldsymbol{X}}^{\otimes r}$ for some $C_r : (\mathbb{R}^n)^{\otimes r} \to \mathbb{R}^p$ satisfying

$\mathrm{span}\{\boldsymbol{u}_i^{\otimes r} : i \in [k]\} \subseteq \mathrm{range}(C_r^\top)$, where $p \le n$. Extend the projected covariate by one tensor degree $(C_r\widetilde{\boldsymbol{X}}^{\otimes r}) \otimes \widetilde{\boldsymbol{X}} \in \mathbb{C}^{pn}$ and form the extended projected Hermite matrix

$$\mathbb{E}_{\boldsymbol{X},\xi,\boldsymbol{X}'}\left[\frac{\mathrm{Re}\left[\left((C_r\widetilde{\boldsymbol{X}}^{\otimes r})\otimes\widetilde{\boldsymbol{X}}\right)\left((C_r\widetilde{\boldsymbol{X}}^{\otimes r})\otimes\widetilde{\boldsymbol{X}}\right)^\top\right]}{\sqrt{(2r+2)!}}\mathbf{1}\{T \le Y_{\max} \le T+1\}\right]$$
$$\approx \sum_{i\in[k]} \lambda_i^{(2r+2)}\left[(C_r\boldsymbol{u}_i^{\otimes r})\otimes\boldsymbol{u}_i\right]\left[(C_r\boldsymbol{u}_i^{\otimes r})\otimes\boldsymbol{u}_i\right]^\top .$$

The $\approx$ here is due to the remainder term R_{2r+2}. Let $V_{r+1} \in \mathbb{R}^{k\times pn}$ be the matrix whose rows are the k leading eigenvectors of the extended projected Hermite matrix on the left (padding with arbitrary orthonormal vectors if necessary). This allows us to extend the degree-r projection map to degree $r+1$ as $C_{r+1} := V_{r+1}(C_r \otimes I_n)$.

Using this technique, we can iteratively extend the projection up to degree s. Note that it suffices to store the matrices $V_2, \ldots, V_s$ with at most $knp \le kn^2$ entries each and we can implicitly evaluate projections of the covariate via

$$C_1\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{X}}\,, \qquad C_r\widetilde{\boldsymbol{X}}^{\otimes r} = V_r\left[(C_{r-1}\widetilde{\boldsymbol{X}}^{\otimes(r-1)})\otimes\widetilde{\boldsymbol{X}}\right]\,, \quad 2 \le r \le s\,.$$

By construction, we always have $\mathrm{span}\{\boldsymbol{u}_i^{\otimes r} : i \in [k]\} \subseteq \mathrm{range}(C_r^\top)$. The first intermediate matrix has size $n^2 \times n^2$, and the later matrices have size at most $kn \times kn$. Hence we can indeed compute (4) in polynomial time.

We have successfully reduced dimensions of the higher-order Hermite tensors to $O(k^2)$. However, the projected vectors $C_r\boldsymbol{u}_i^{\otimes r}$ do not directly reveal the original regressors. Next, we will see how to circumvent this issue using odd-degree Hermite tensors in addition to the even-degree Hermite tensors we have seen so far. This requires borrowing some tools from the observable-operator learning literature.

2.2.3 Observable-Operator Learning

Abstractly speaking, observable-operator learning is a technique for learning latent-variable models by constructing operators from samples whose eigenvectors or eigenvalues encode the latent variables [LRA93; AHK12]. We will design a way to encode the regressors in the spectrum of the projected Hermite tensors. This allows us to apply existing tools from the observable-operator learning literature due to Hsu and Kakade [HK13] for learning spherical GMMs. Our algorithms and exposition found in Section 6.3 follow [HK13] but the analysis is our own, as we need to carefully handle the irreducible remainder terms R_{2s} and R_{2s+1}.

Reduction from Projected Hermite Tensors. Assume we are able to efficiently compute projections of the complexified covariates $C_s\widetilde{\boldsymbol{X}}^{\otimes s}$ as above. This enables us to compute projections of the flattened degree-$2s$ Hermite matrices

$$C_s\mathbb{E}_{\boldsymbol{X},\xi}\left[\mathrm{Mat}_s(\mathsf{H}_{2s}(\boldsymbol{X}))\mathbf{1}\{T \le Y_{\max} \le T+1\}\right]C_s^\top \approx \underbrace{\sum_{i\in[k]} \lambda_i^{(2s)}(C_s\boldsymbol{u}_i^{\otimes s})(C_s\boldsymbol{u}_i^{\otimes s})^\top}_{=:A}\,.$$

We can similarly consider odd-degree Hermite tensors where the last tensor component is contracted with some vector $\boldsymbol{b} \in \mathbb{R}^n$. As it turns out, for some $\boldsymbol{a}_i := \alpha_i \boldsymbol{w}_i$ where $\alpha_i > 0$ is a positive scalar,

$$\begin{aligned}
&C_s \, \mathbb{E}_{\boldsymbol{X},\xi} \left[\mathrm{Mat}_s(\mathsf{H}_{2s+1}(\boldsymbol{X}) \lrcorner \boldsymbol{b}) \mathbf{1}\{T \le Y_{\max} \le T+1\} \right] C_s^\top \\
&\quad = \mathbb{E}_{\boldsymbol{X},\xi,\boldsymbol{X}'} \left[\frac{\mathrm{Re}[C_s \widetilde{\boldsymbol{X}}^{\otimes s} (C_s \widetilde{\boldsymbol{X}}^{\otimes s})^\top \left\langle \boldsymbol{b}, \widetilde{\boldsymbol{X}} \right\rangle]}{\sqrt{(2s+1)!}} \mathbf{1}\{T \le Y_{\max} \le T+1\} \right] \\
&\quad \approx \underbrace{\sum_{i \in [k]} \lambda_i^{(2s)} \langle \boldsymbol{b}, \boldsymbol{a}_i \rangle \left[(C_s \boldsymbol{u}_i^{\otimes s}) \right] \left[(C_s \boldsymbol{u}_i^{\otimes s}) \right]^\top}_{=: M(\boldsymbol{b})} .
\end{aligned}$$

Here the $\approx$ is only due to the remainder term R_{2s+1}. Note that the projected even and odd matrices A and $M(\boldsymbol{b})$ differ only in the spectrum and that $M(\boldsymbol{b})$ encodes a scaled version of the regressors within its spectrum. Moreover, the vector $\boldsymbol{b} \in \mathbb{R}^n$ was arbitrary. Hence we can hope to recover the scaled regressors $\boldsymbol{a}_i$ using A and several instantiations of $M(\boldsymbol{b})$!

Observable-Operator Learning. Compute $A^{-1/2}$ from an eigendecomposition of A and remove the component weights from $M(\boldsymbol{b})$:

$$D(\boldsymbol{b}) := A^{-\frac{1}{2}} M(\boldsymbol{b}) A^{-\frac{1}{2}} = \sum_{i \in [k]} \langle \boldsymbol{b}, \boldsymbol{a}_i \rangle \, \boldsymbol{v}_i \boldsymbol{v}_i^\top .$$

The collection $\{D(\boldsymbol{b})\}_{\boldsymbol{b} \in \mathbb{R}^n}$ is sometimes known as an *observable-operator family*. For a random unit vector $\boldsymbol{b}$, the eigenvalues $\langle \boldsymbol{b}, \boldsymbol{a}_i \rangle$ are distinct with probability one, so an eigendecomposition of $D(\boldsymbol{b})$ recovers the common eigenvectors $\boldsymbol{v}_i$. Then, by evaluating $D(\boldsymbol{b})$ at the standard basis vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$, we can extract the scaled regressors $\boldsymbol{a}_i$ using the eigenvectors. That is, the j-th coordinate of $\boldsymbol{a}_i$ can be computed as $(\boldsymbol{a}_i)_j = \boldsymbol{v}_i^\top D(\boldsymbol{e}_j) \boldsymbol{v}_i$.

Additional Complications. This leaves us with two additional complications. The first is that we have only computed some positive scalings of the regressors $\boldsymbol{a}_i = \alpha_i \boldsymbol{w}_i$ and need to somehow recover the lengths of the regressors. In Section 2.2.4, we will explain a simple refinement procedure that recovers a regressor $\boldsymbol{w}_i$ given a unit-norm estimate $\widehat{\boldsymbol{u}}_i \approx \boldsymbol{u}_i = \boldsymbol{w}_i / \|\boldsymbol{w}_i\|_2$. We can then run this procedure with $\widehat{\boldsymbol{u}}_i = \boldsymbol{a}_i / \|\boldsymbol{a}_i\|_2$.

The second more subtle issue is that the magnitude of the remainder term R_{2s} can exceed the minimum eigenvalue $\min_i \lambda_i^{(2s)}$ of A, so that inverting estimates of A to estimate its inverse square root is an ill-conditioned operation and may exponentially amplify estimation error. This can be handled by truncating the spectrum of A at its largest spectral gap to obtain $\bar{A}$. This ensures that inverting $\bar{A}$ will not blow up the estimation error. However, carrying through the observable-operator learning and subsequent refinement procedures using $\bar{A}$ in place of A only recovers a subset of the regressors.

This still leaves some unrecovered regressors. In Section 2.2.5, we explain how to simulate the Hermite moments of the reduced self-selection problem induced by the unrecovered regressors using only observations from the original problem as well as the recovered regressors. Thus we can simply recurse on this smaller instance and eventually recover all regressors.

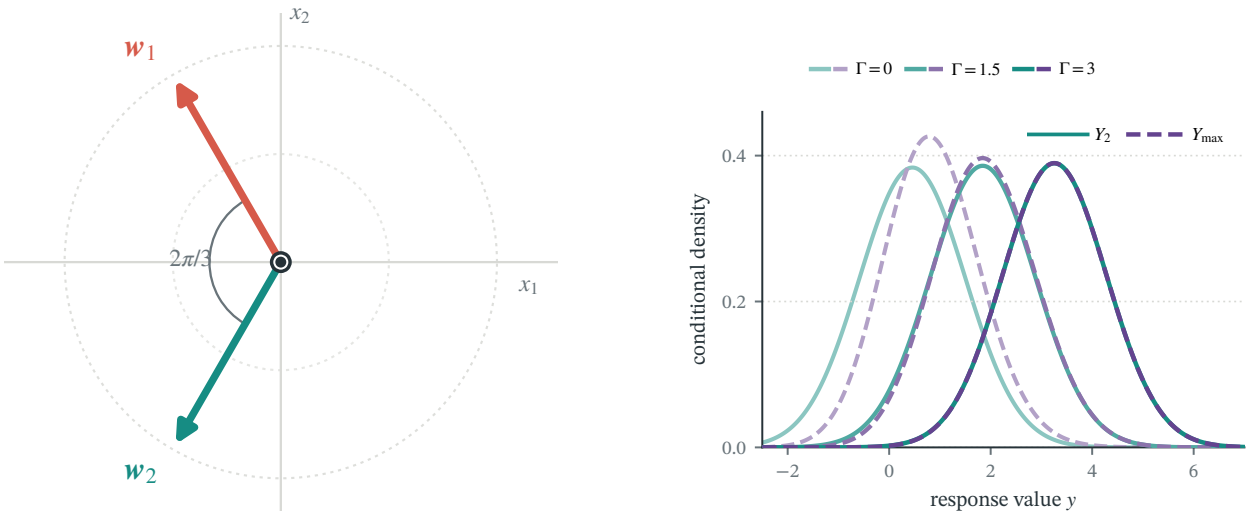
2.2.4 Restricted Least-Squares

Consider the classical linear regression setting, where n-dimensional covariates $\boldsymbol{X} \sim \mathcal{D}$ are sampled from an arbitrary distribution and noisy responses $Y = \boldsymbol{w}^\top \boldsymbol{X} + \xi$ are observed, with independent noise $\xi \sim N(0,1)$. As long as the covariates have an invertible second moment, $\mathbb{E}_{\boldsymbol{X}}[\boldsymbol{X}\boldsymbol{X}^\top] \succ 0$, then we can recover $\boldsymbol{w}$ via ordinary least squares (OLS):

$$\boldsymbol{w} = \mathbb{E}_{\boldsymbol{X}}[\boldsymbol{X}\boldsymbol{X}^\top]^{-1}\mathbb{E}_{\boldsymbol{X}}[\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{w}] = \mathbb{E}_{\boldsymbol{X}}[\boldsymbol{X}\boldsymbol{X}^\top]^{-1}\mathbb{E}_{\boldsymbol{X},\xi}[\boldsymbol{X}Y]\,. \tag{6}$$

Below, we will explain the key steps to reduce the problem of recovering $\boldsymbol{w}_i$ to OLS, given only directional estimates of the regressors, and defer further details to Section 7.

Reduction to Restricted Least-Squares. Let us return to the self-selection setting and assume now that we have an estimate $\widehat{\boldsymbol{u}}_i \approx \boldsymbol{u}_i = \boldsymbol{w}_i / \|\boldsymbol{w}_i\|_2$ for some $i \in [k]$. Our goal is to estimate $\boldsymbol{w}_i$ up to some desired error ε_{int}. Note that we do not need to know the precise regressor index i, just that we have a good estimate to some normalized regressor. As the discussion for ordinary least-squares above holds for an arbitrary distribution $\mathcal{D}$, our strategy is to restrict to an event $\mathcal{E}_{\widehat{\boldsymbol{u}}_i}$ on which $Y_{\max} = Y_i$ with high probability. Then, we can simply use a plug-in estimator for (6) with $\boldsymbol{X}\boldsymbol{X}^\top \mathbf{1}_{\mathcal{E}_{\widehat{\boldsymbol{u}}_i}}$ and $\boldsymbol{X}Y_{\max}\mathbf{1}_{\mathcal{E}_{\widehat{\boldsymbol{u}}_i}}$.



(a) Regressors

(b) Conditional response densities

Figure 2: For the regressors in (a), conditioned on $\Gamma \le \langle \boldsymbol{u}_2, \boldsymbol{X} \rangle \le \Gamma + 1$, the density of the maximum response $Y_{\max}$ becomes indistinguishable from that of Y_2 as Γ grows (b).

Gaitonde and Mossel [GM24] showed that if a candidate is sufficiently close to $\boldsymbol{w}_i$, then Y_i attains the maximum with high probability conditional on a large projection along the candidate direction. We use the same idea with the event

$$\mathcal{E}_{\widehat{\boldsymbol{u}}_i} := \{\Gamma \le \langle \widehat{\boldsymbol{u}}_i, \boldsymbol{X} \rangle \le \Gamma + 1\}\,.$$

Intuitively, since the regressors have constant angular separation, if a covariate is highly correlated with a single regressor $\boldsymbol{w}_i$, it cannot also be highly correlated with other regressors $\boldsymbol{w}_j, j \neq i$, so

the maximizer response should be Y_i with high probability. To see this, we first remark that the separation condition (Assumption 1.2) ensures that $\langle \boldsymbol{w}_i - \boldsymbol{w}_j, \widehat{\boldsymbol{u}}_i \rangle = \frac{\|\boldsymbol{w}_i\|_2^2 - \langle \boldsymbol{w}_i, \boldsymbol{w}_j \rangle}{\|\boldsymbol{w}_i\|_2} \geq \Omega(1)$. Hence the conditional distribution satisfies

$$(Y_i - Y_j \mid \langle \widehat{\boldsymbol{u}}, \boldsymbol{X} \rangle = \Gamma) \sim N(\mu, \sigma^2), \qquad \mu \geq \Omega(\Gamma), \sigma^2 = O(1).$$

We can thus choose $\Gamma = \Theta(\sqrt{\log(n/\varepsilon_{\text{int}})})$ so that by a union bound, the probability that Y_i does not attain the max is at most $ke^{-\Omega(\Gamma^2)} = \text{poly}(\varepsilon_{\text{int}}/n)$.

Local Convergence. As a corollary of our restricted least squares algorithm, we get a simple polynomial-time local convergence algorithm for self-selection improving on prior works. Kalavasis, Mehrotra, and Zhou [KMZ26] designed a polynomial-time local convergence algorithm for self-selection that required estimates $\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2 \leq \text{poly}(1/k)$ with inverse polynomial initialization distance. We note that $\|\widehat{\boldsymbol{w}}_i - \boldsymbol{w}_i\|_2 = O(1)$ implies that $\|\widehat{\boldsymbol{w}}_i / \|\widehat{\boldsymbol{w}}_i\|_2 - \boldsymbol{u}_i\|_2 = O(1)$, possibly with bigger constants. Hence our restricted least squares algorithm immediately implies a local convergence algorithm with $O(1)$ convergence radius, improving on [KMZ26].

2.2.5 Moment Peeling

As previously mentioned, our observable-operator learning subroutine extracts estimates for a block of the normalized regressors which can be refined to actual regressor estimates using restricted least squares. The final piece of the puzzle is to show that we can recursively run our algorithm to recover the rest of the regressors. This is not an issue for the restricted least squares algorithm as it only requires an accurate angular estimate of some regressor. For our implicit tensor decomposition subroutine, our key observation is that we only need to access samples when computing the observable-operator family. In particular, we only need to estimate restricted Hermite tensors of the covariate. Below, we describe the intuition for how to simulate these tensors for the unrecovered regressors which is all we need to run our algorithm and defer further details to Section 8. This idea of iteratively recovering and removing components has been used in self-selection [CDIZ23], Gaussian mixture clustering [LL24], and general tensor decomposition [AGHK+14].

Moment Peeling. Let $R \subseteq [k], Q := [k] \setminus R$ and assume we recovered $\boldsymbol{w}_i, i \in R$. Our goal is to be able to simulate restricted Hermite tensors for the (unobserved) maximum $Y_{\max,Q} = \max_{i \in Q} Y_i$,

$$\mathbb{E}_{\boldsymbol{X},\xi}\left[\mathsf{H}_d(\boldsymbol{X})\mathbf{1}\{T \leq Y_{\max,Q} \leq T+1\}\right].$$

We construct a weight function $\omega_R(\boldsymbol{X}, Y_{\max})$ whose conditional expectation approximates that of $\mathbf{1}\{T \leq Y_{\max,Q} \leq T+1\}$,

$$\mathbb{E}_{\xi}[\omega_R(\boldsymbol{X}, Y_{\max}) \mid \boldsymbol{X}] \approx \mathbb{E}_{\xi}[\mathbf{1}\{T \leq Y_{\max,Q} \leq T+1\} \mid \boldsymbol{X}]$$

using only the observations and the recovered regressors. We emphasize that we do not observe $Y_{\max,Q}$ directly.

Let $Y_{\max,R} := \max_{i\in R} Y_i$ and note that $Y_{\max} = \max(Y_{\max,R}, Y_{\max,Q})$. We begin with the pointwise identity

$$\begin{aligned}
&\mathbf{1}\{T \le Y_{\max} \le T+1\} + \mathbf{1}\{T \le \min(Y_{\max,R}, Y_{\max,Q}) \le T+1\} \\
&\quad = \mathbf{1}\{T \le Y_{\max,R} \le T+1\} + \mathbf{1}\{T \le Y_{\max,Q} \le T+1\}\,,
\end{aligned}$$

$$\begin{aligned}
&\mathbf{1}\{T \le Y_{\max} \le T+1\} - \mathbf{1}\{T \le Y_{\max,R} \le T+1\} \\
&\quad = \mathbf{1}\{T \le Y_{\max,Q} \le T+1\} - \mathbf{1}\{T \le \min(Y_{\max,R}, Y_{\max,Q}) \le T+1\}\,. \qquad (7)
\end{aligned}$$

Now, the RHS is nearly the indicator variable we want modulo a bias term. Intuitively, this bias term is a much rarer event since it roughly requires at least two Gaussians to exceed T, while all other events roughly only require one.

We would like to define $\omega_R(\boldsymbol{X}, Y_{\max})$ to be the LHS. Although $\mathbf{1}\{T \le Y_{\max,R} \le T+1\}$ is unobserved, we only need the weight function to be a good approximation under conditional expectation. Recall Φ is the CDF of the standard normal distribution. It suffices to take

$$\begin{aligned}
\omega_R(\boldsymbol{X}, Y_{\max}) &:= \mathbf{1}\{T \le Y_{\max} \le T+1\} - \left(\prod_{i\in R} \Phi(T+1-\boldsymbol{w}_i^\top \boldsymbol{X}) - \prod_{i\in R} \Phi(T-\boldsymbol{w}_i^\top \boldsymbol{X})\right) \\
\mathbb{E}_\xi[\omega_R(\boldsymbol{X}, Y_{\max}) \mid \boldsymbol{X}] &= \mathbb{E}_\xi\left[\mathbf{1}\{T \le Y_{\max} \le T+1\} - \mathbf{1}\{T \le Y_{\max,R} \le T+1\} \mid \boldsymbol{X}\right].
\end{aligned}$$

Now, introducing the Hermite tensor $\mathsf{H}_d(\boldsymbol{X})$, taking the expectation over the randomness $(\boldsymbol{X}, \xi)$, and using the identity (7), we have that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X},\xi}\left[\mathsf{H}_d(\boldsymbol{X})\omega_R(\boldsymbol{X}, Y_{\max})\right] \\
&\quad = \mathbb{E}_{\boldsymbol{X},\xi}\left[\mathsf{H}_d(\boldsymbol{X})\mathbf{1}\{T \le Y_{\max,Q} \le T+1\}\right] - \mathbb{E}_{\boldsymbol{X},\xi}\left[\mathsf{H}_d(\boldsymbol{X})\mathbf{1}\{T \le \min(Y_{\max,R}, Y_{\max,Q}) \le T+1\}\right].
\end{aligned}$$

Thus we can indeed produce estimates of the desired restricted Hermite moment up to some bias term.

This concludes the overview of our techniques.

2.3 Organization

We organize the rest of this work as follows. Related work is discussed in Section 3 and some preliminaries are collected in Section 4. In Section 5, we show that high-order restricted Hermite tensors approximately decompose as a weighted sum of near-orthogonal rank-one tensors. In Section 6, we use this structure to design an implicit tensor decomposition procedure that recovers a subset of normalized regressors without explicitly forming the moment tensors. Section 7 uses restricted least squares to estimate the full regressors from these unit vector estimates. Section 8 implements an important estimation subroutine used in our implicit tensor decomposition procedure. We combine these steps in Section 9 to prove Theorem 1.3.

For convenience, we indicate the main dependencies in the proof of Theorem 1.3 in Figure 3.

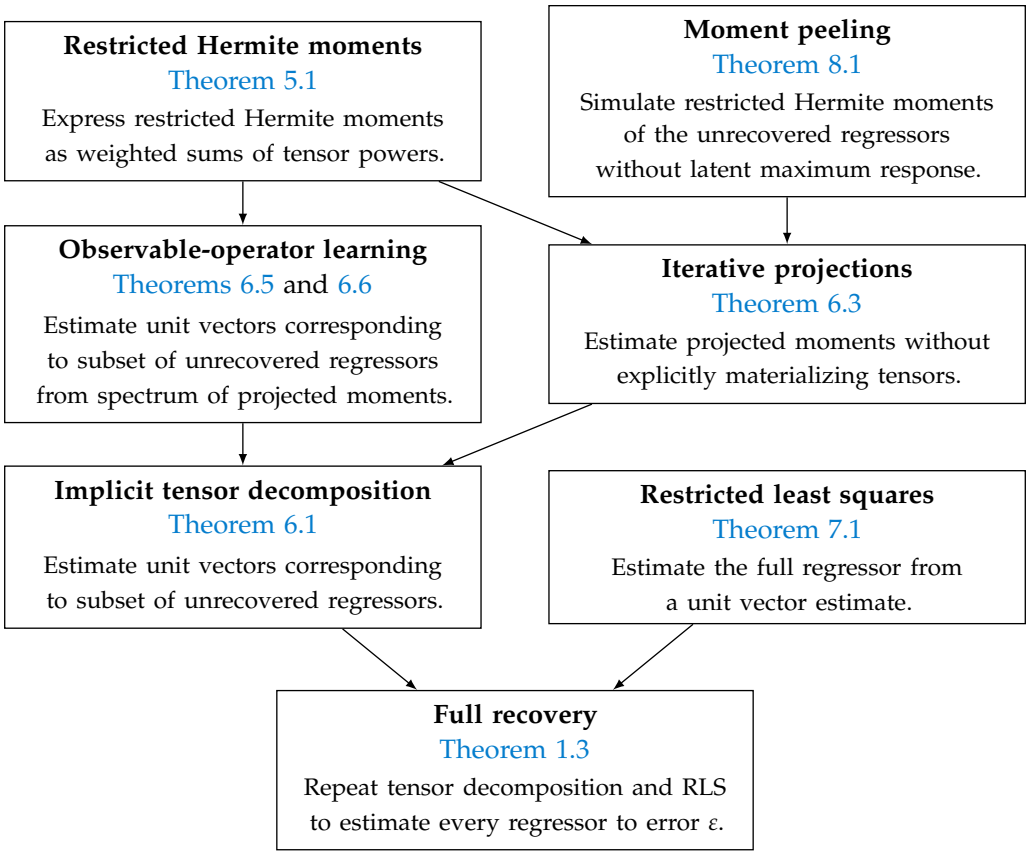


Figure 3: Main dependencies in the proof of Theorem 1.3. Each arrow points from a result to a downstream application.

3 Related Works

Regression with Self-Selection. As discussed, Cherapanamjeri, Daskalakis, Ilyas, and Zampetakis [CDIZ23] initiated the study of time- and sample-efficient algorithms for regression with self-selection. Subsequent work improved the sample complexity and running time, but the best known end-to-end algorithms still require time exponential in the number of regressors k [GM24; KMZ26]. Our main contribution is the first polynomial-time algorithm for regression with self-selection.

Max-Affine Regression. Max-affine regression estimates a maximum of affine functions from noisy observations. Estimation algorithms and statistical guarantees have been studied for both general and sparse models [MB09; BGS15; GPGR22; KBL24; KL24; KKL26]. The main difference from the self-selection model is that the noise is added to the response after taking the maximum, so the noise does not affect the selection of the affine function.

Gaussian Mixture Models. Gaussian mixture learning seeks to estimate a mixture's density or parameters, or to cluster its samples. Learning and clustering guarantees have been studied under various structural assumptions and in the presence of corrupted samples [Das99; VW04; AM05;

AK05; BS10; KMV10; MV10; HK13; GHK15; RV17; HL18; KSS18; BDJK+22; LM22]. Our iterative projection construction follows work on clustering separated spherical Gaussian mixtures using implicit high-order Hermite moments [LL24].

Mixtures of Linear Regressions. In mixed linear regression, each response is generated by an unobserved regressor sampled independently of the covariate. Parameter recovery in linear and generalized-linear mixtures has been studied under different assumptions on the covariates, noise, and initialization [De 89; CL13; YCS14; SJA16; ZJD16; BWY17; LL18; CLS20; DK20; KC20]. More directly related to our approach, implicit high-order moments yield polynomial-time algorithms for density estimation [DK25].

Multi-View Mixtures and Hidden Markov Models. In multi-view mixtures, conditionally independent observations share a latent component; in hidden Markov models, the latent states form a Markov chain. Both admit learning algorithms under suitable rank and identifiability assumptions [Jae00; HKZ12; AGHK+14]. Our observable-operator family uses the same common-eigenbasis construction developed for multi-view mixtures [AHK12].

Gaussian Multi-Index Models. In Gaussian multi-index models, the response depends on the covariate through an unknown low-dimensional projection. Algorithms and statistical limitations have been studied for recovering this subspace or learning the response function [Li91; Li92; Vem10; CM20; DPLB24]. Our recursive projection construction is related to sequential recovery of directions from Hermite or conditional moments, although recovering the subspace does not identify the individual regressors in our model [DLB25; DIKR25; DIKZ25].

Learning Shallow Neural Networks. Learning a shallow network asks to learn a sum of activations of linear forms or recover its weights from labeled examples. Learning guarantees under Gaussian covariates have been studied for several activations and network classes [JSA15; ZSJB+17; DKKZ20; DLS22; CN24]. Our use of random contractions and comparisons between Hermite degrees is related to algorithms for learning ReLU networks [CDGK+23; DK24].

Learning from Coarse Observations. Coarse observations reveal only a set containing the latent sample. Results include learning algorithms for coarse labels and Gaussian mean estimation under identifiable convex partitions [FKKT21; KMZZ+26]. This general framework captures the self-selection problem and through this connection, local convergence algorithms for self-selection have been developed [KMZ26].

Truncated and Censored Data. In truncated or censored estimation, observations are either discarded or only partially revealed [Tob58]. Algorithms for known truncation estimate regression and Gaussian parameters under suitable survival and design assumptions, with applications to differentially private estimation [DGTZ18; DGTZ19; DRZ20; IZD20; ZZ25]. Unknown truncation has also been studied for regression and exponential-family estimation under structural assumptions on the survival set and distribution [KTZ19; LMZ24; KMKC26; LMZ26; LSSW26]. Other related work

studies general selection filters, computational lower bounds, and truncation detection [DLNS24; DKPZ24; KLMZ26].

4 Preliminaries

In this section, we introduce the notational conventions as well as some preliminary tools used throughout this work.

4.1 Hyperparameters

We choose the tensor order s to make the lifted vectors nearly orthogonal, and the tail threshold T so that the decay of the tail probability dominates the growth of the Hermite moments. The internal accuracy $\varepsilon_{\mathrm{int}}$ is chosen to be sufficiently small that errors that accumulate across rounds can be absorbed. The refinement threshold Γ is chosen to control the bias of the restricted least squares estimate.

Hyperparameters 4.1. Throughout this work, we fix the following hyperparameters

$$K := \frac{16(1+B^2)^3}{\Delta^4}, \tag{8}$$

$$s := K^4 \log n, \tag{9}$$

$$T := 2K^2\sqrt{s}, \tag{10}$$

$$\varepsilon_{\mathrm{int}} := \min\{\varepsilon, e^{-K^6 s}\}, \tag{11}$$

$$\Gamma := 2K\sqrt{\log\left(\frac{n}{\varepsilon_{\mathrm{int}}}\right)}. \tag{12}$$

4.2 Notation

For a positive integer m, write $[m] = \{1, \dots, m\}$. We use bold letters for vectors, ordinary capital letters for matrices, and sans-serif capital letters for tensors. We use $\langle \cdot, \cdot \rangle$ for the Euclidean inner product on vectors and the Frobenius inner product on real matrices and tensors, and $\|\cdot\|_2, \|\cdot\|_F, \|\cdot\|_{\mathrm{op}}$ for the Euclidean, Frobenius, and operator norms. The symbol $\mathbf{1}\{\mathcal{E}\}$ denotes the indicator of an event $\mathcal{E}$.

We use $\otimes$ for tensor products and $\boldsymbol{v}^{\otimes d}$ for the d-fold tensor product of $\boldsymbol{v}$. The space $\mathrm{Sym}^d(\mathbb{R}^n)$ consists of symmetric order-d tensors on $\mathbb{R}^n$. The notation $\mathsf{A} \lrcorner \boldsymbol{b}$ denotes contraction against $\boldsymbol{b}$ in the last tensor factor. For an order-$2r$ tensor on $\mathbb{R}^n$, $\mathrm{Mat}_r(\cdot)$ flattens its argument into an $n^r \times n^r$ matrix using the first r multi-indices as the row index and the last r as the column index. We write $\mathrm{i} = \sqrt{-1}$ and use Re for the entrywise real part. Finally, Φ denotes the standard Gaussian CDF and ϕ_{σ^2} the density of $N(0, \sigma^2)$. We write $\phi = \phi_1$ for the standard Gaussian density.

4.3 Properties of Gaussians

Proposition 4.2 (Gaussian Tail Bound; [Wai19, Equation (2.7)]). *Let $G \sim N(\mu, \sigma^2)$ with $\sigma > 0$. For every $t \geq 0$,*

$$\mathbb{P}_G(G \geq \mu + t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right).$$

By symmetry, the same bound holds for $\mathbb{P}_G(G \leq \mu - t)$.

Proposition 4.3 (Conditional Gaussian Distribution; [RW06, Equation (A.6)]). *Let $\boldsymbol{G} \in \mathbb{R}^n$ and $Y \in \mathbb{R}$ be centered jointly Gaussian. Write $\Sigma = \mathbb{E}_{\boldsymbol{G}}[\boldsymbol{G}\boldsymbol{G}^\top]$, $\boldsymbol{v} = \mathbb{E}_{\boldsymbol{G},Y}[\boldsymbol{G}Y]$, and $\sigma^2 = \mathbb{E}_Y[Y^2] > 0$. For every $t \in \mathbb{R}$,*

$$\boldsymbol{G} \mid Y = t \sim N\left(\frac{\boldsymbol{v}t}{\sigma^2}, \Sigma - \frac{\boldsymbol{v}\boldsymbol{v}^\top}{\sigma^2}\right).$$

Proposition 4.4 (Gaussian Hypercontractivity for Polynomials; [ODo14, Theorem 11.23 and Proposition 11.37]). *Fix integers $n \geq 1$ and $d \geq 0$, and a real polynomial $P : \mathbb{R}^n \to \mathbb{R}$ of total degree at most d. Let $\boldsymbol{G} \sim N(\boldsymbol{0}, I_n)$. For every $2 \leq p < \infty$,*

$$\|P\|_{L^p(\boldsymbol{G})} \leq (p-1)^{d/2} \|P\|_{L^2(\boldsymbol{G})}, \qquad \|P\|_{L^p(\boldsymbol{G})} := \left(\mathbb{E}_{\boldsymbol{G}}[|P(\boldsymbol{G})|^p]\right)^{1/p}.$$

4.4 Hermite Polynomials

We define the Hermite features and fix their normalization. Hermite expansions have been used in algorithms and statistical query lower bounds for learning neural networks under Gaussian inputs [DKKZ20; DK24].

Let $\boldsymbol{X}' \sim N(\boldsymbol{0}, I_n)$ be an auxiliary Gaussian vector.

For every nonnegative integer d, the probabilists' Hermite polynomial has the Gaussian integral representation [DLMF, Equations 18.10.7 and 18.7.11]

$$\mathrm{He}_d(t) := \mathbb{E}_g[(t + \mathrm{i}g)^d], \qquad g \sim N(0,1), \quad t \in \mathbb{R}.$$

This expectation is real, since g and $-g$ have the same law. For $\boldsymbol{x} \in \mathbb{R}^n$, define the normalized Hermite tensor

$$\mathsf{H}_d(\boldsymbol{x}) := \frac{1}{\sqrt{d!}} \mathbb{E}_{\boldsymbol{X}'}[\mathrm{Re}((\boldsymbol{x} + \mathrm{i}\boldsymbol{X}')^{\otimes d})]. \tag{13}$$

These tensors are real and symmetric. For example,

$$\mathsf{H}_0(\boldsymbol{x}) = 1, \qquad \mathsf{H}_1(\boldsymbol{x}) = \boldsymbol{x}, \qquad \mathsf{H}_2(\boldsymbol{x}) = \frac{\boldsymbol{x}\boldsymbol{x}^\top - I_n}{\sqrt{2}}.$$

In one dimension, $\mathsf{H}_d(t) = \mathrm{He}_d(t)/\sqrt{d!}$.

The normalized Hermite polynomials are orthonormal under a standard Gaussian [ODo14, Proposition 11.33]. For $G \sim N(0,1)$ and nonnegative integers d_1, d_2,

$$\mathbb{E}_G\left[\frac{\mathrm{He}_{d_1}(G)}{\sqrt{d_1!}} \frac{\mathrm{He}_{d_2}(G)}{\sqrt{d_2!}}\right] = \begin{cases} 1, & d_1 = d_2, \\ 0, & d_1 \neq d_2. \end{cases} \tag{14}$$

Thus Hermite polynomials of different degrees are uncorrelated when evaluated at a standard Gaussian.

For a symmetric order-d tensor A, its entries are unchanged when the indices are permuted. We group these entries by the number of times each coordinate appears. Write $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ for these counts, and let $A_{\boldsymbol{\alpha}}$ denote their common value. For example, for an order-three tensor on $\mathbb{R}^2$, $\boldsymbol{\alpha} = (2, 1)$ groups the entries $A_{112} = A_{121} = A_{211}$. Write $|\boldsymbol{\alpha}| = \sum_j \alpha_j$ and $\boldsymbol{\alpha}! = \prod_j \alpha_j!$. There are $d!/\boldsymbol{\alpha}!$ entries in the group indexed by $\boldsymbol{\alpha}$.

Proposition 4.5 (Hermite Expansion and Normalization; [ODo14, Proposition 11.33])**.** *Let* A *be a symmetric order-d tensor. For every $\boldsymbol{x} \in \mathbb{R}^n$,*

$$\langle \mathsf{H}_d(\boldsymbol{x}), \mathsf{A} \rangle = \sum_{|\boldsymbol{\alpha}|=d} \sqrt{\frac{d!}{\boldsymbol{\alpha}!}}\, A_{\boldsymbol{\alpha}} \prod_{j=1}^{n} \frac{\mathrm{He}_{\alpha_j}(x_j)}{\sqrt{\alpha_j!}}.$$

For $\boldsymbol{G} \sim N(\mathbf{0}, I_n)$, the product form of Hermite orthogonality shows that the products in this sum are orthonormal. Consequently,

$$\mathbb{E}_{\boldsymbol{G}}[\langle \mathsf{H}_d(\boldsymbol{G}), \mathsf{A} \rangle^2] = \sum_{|\boldsymbol{\alpha}|=d} \frac{d!}{\boldsymbol{\alpha}!} A_{\boldsymbol{\alpha}}^2 = \|\mathsf{A}\|_F^2 . \tag{15}$$

We give the derivation in Appendix B.1.

4.5 Spectral Perturbation Bounds

We use Weyl's eigenvalue perturbation inequality in the following form.

Proposition 4.6 (Weyl's Eigenvalue Perturbation Inequality; [Bha97, Corollary III.2.6])**.** *Let $H, \widetilde{H} \in \mathbb{R}^{p \times p}$ be symmetric, with eigenvalues in nonincreasing order. Then, for every $j \in [p]$,*

$$\left|\lambda_j(\widetilde{H}) - \lambda_j(H)\right| \le \left\|\widetilde{H} - H\right\|_{\mathrm{op}} .$$

4.6 Heavy-Tail Mean Estimation

We write $\textsc{TrimmedMean}(\boldsymbol{W}; \mathsf{v}, \varepsilon, \delta)$ for the coordinatewise trimmed-mean estimator with sample access to $\boldsymbol{W}$, coordinate second-moment bound v, coordinate accuracy ε, and failure probability δ. The estimator draws independent copies of $\boldsymbol{W}$ and averages each coordinate after removing its smallest and largest observations. For matrices or families of matrices, we apply the estimator entrywise using the same observations throughout. The following consequence of [OOR25, Theorem 1.1.1] gives its guarantee.

Lemma 4.7 (Coordinatewise Trimmed Mean)**.** *Let $\boldsymbol{W} = (W_1, \ldots, W_p)$ be a random vector satisfying*

$$\mathbb{E}_{\boldsymbol{W}}[W_j^2] \le \mathsf{v} \qquad (j \in [p]),$$

where $\mathsf{v} > 0$. For $0 < \varepsilon \le \sqrt{\mathsf{v}}$ and $\delta \in (0, 1/3)$, $\textsc{TrimmedMean}(\boldsymbol{W}; \mathsf{v}, \varepsilon, \delta)$ returns $\widehat{\boldsymbol{\mu}} \in \mathbb{R}^p$ satisfying

$$\max_{j \in [p]} |\widehat{\mu}_j - \mathbb{E}_{\boldsymbol{W}}[W_j]| \le \varepsilon$$

with probability at least $1 - \delta$, using $O(\mathsf{v}\varepsilon^{-2} \log(p/\delta))$ independent copies of $\boldsymbol{W}$.

The construction and proof are given in Appendix B.2.

5 Structure of Restricted Hermite Moments

In this section, we show that the restricted Hermite moments of the latent maximum corresponding to unrecovered regressors are close to weighted sums of tensor powers of said regressors. As we will see, the even and odd moments have the same rank-one components, and the odd moments allow us to recover the regressor directions. We also establish the bounds needed to estimate and project these moments and to apply the observable-operator learning procedure.

Fix a non-empty set $Q \subseteq [k]$ of remaining regressor indices. Recall that $Y_i = \langle \boldsymbol{w}_i, \boldsymbol{X} \rangle + \xi_i$ is the latent response of component i and $Y_{\max,Q} = \max_{i \in Q} Y_i$ is the remaining maximum. Write $\sigma_i^2 = 1 + \|\boldsymbol{w}_i\|_2^2$ for the variance of Y_i and $\boldsymbol{u}_i = \boldsymbol{w}_i / \|\boldsymbol{w}_i\|_2$ for its normalized regressor.

Restricted Hermite Moments. We use the normalized Hermite tensors from Section 4.4. Let $\boldsymbol{X}' \sim N(\boldsymbol{0}, I_n)$ be independent of the covariate and latent noises, and put

$$\widetilde{\boldsymbol{X}} := \boldsymbol{X} + \mathrm{i}\boldsymbol{X}'.$$

Our algorithm uses this complexified covariate to evaluate the Hermite features.

We consider the restricted Hermite moments

$$\begin{aligned} \mathsf{T}_d^Q &:= \mathbb{E}_{\boldsymbol{X},\xi_Q}[\mathsf{H}_d(\boldsymbol{X})\mathbf{1}\{T \le Y_{\max,Q} \le T+1\}] \\ &= \frac{1}{\sqrt{d!}}\,\mathbb{E}_{\boldsymbol{X},\xi_Q,\boldsymbol{X}'}[\mathrm{Re}(\widetilde{\boldsymbol{X}}^{\otimes d})\mathbf{1}\{T \le Y_{\max,Q} \le T+1\}]\,. \end{aligned} \tag{16}$$

We note that compared to conditional moments, these restricted moments are unnormalized and each expectation is restricted to the tail slice of $Y_{\max,Q}$.

Weights and Scaled Regressors. As we will soon see, the even restricted moments approximately decompose as a weighted sum of rank-one tensors. The odd moments contain the same components, with an additional factor in the weights. We now define these quantities. For $\sigma > 0$, let

$$\phi_{\sigma^2}(t) := \frac{1}{\sqrt{2\pi}\sigma} e^{-t^2/(2\sigma^2)}$$

denote the density of $N(0, \sigma^2)$. Recall $\mathrm{He}_d(\cdot)$ denotes the d-th Hermite polynomial in 1-D. For every integer $d \ge 0$, define the restricted moment coefficient

$$a_d(\sigma^2) := \int_T^{T+1} \phi_{\sigma^2}(t) \frac{\mathrm{He}_d(t/\sigma)}{\sigma^d \sqrt{d!}}\,\mathrm{d}t. \tag{17}$$

For $i \in [k]$ and $1 \le r \le s$, define the degree-$2r$ weights

$$\lambda_i^{(2r)} := a_{2r}(\sigma_i^2)\,\|\boldsymbol{w}_i\|_2^{2r}\,. \tag{18}$$

At level s, define the scaled regressors

$$\boldsymbol{a}_i := \frac{a_{2s+1}(\sigma_i^2)}{a_{2s}(\sigma_i^2)}\,\boldsymbol{w}_i. \tag{19}$$

We are now ready to state the main result of this section.

Theorem 5.1 (Restricted Moment Structure). *Consider the Gaussian self-selection model (Definition 1.1) under the boundedness and uncoveredness assumptions (Assumption 1.2). Use the hyperparameters K, T, s from Hyperparameters 4.1. The following holds for every non-empty $Q \subseteq [k]$.*

(i) For every $1 \le r \le s$ and $\boldsymbol{b} \in \mathbb{R}^n$,

$$\mathrm{Mat}_r(\mathsf{T}^Q_{2r}) = \sum_{i \in Q} \lambda_i^{(2r)} \boldsymbol{u}_i^{\otimes r} (\boldsymbol{u}_i^{\otimes r})^\top + E_r \,, \tag{20}$$

$$\mathrm{Mat}_s(\mathsf{T}^Q_{2s+1} \lrcorner \boldsymbol{b}) = \sum_{i \in Q} \lambda_i^{(2s)} \langle \boldsymbol{b}, \boldsymbol{a}_i \rangle \, \boldsymbol{u}_i^{\otimes s} (\boldsymbol{u}_i^{\otimes s})^\top + E(\boldsymbol{b}) \,. \tag{21}$$

The errors satisfy

$$\begin{aligned} \|E_r\|_F &\le e^{8s \log K - T^2/(2K)} \sum_{i \in Q} \lambda_i^{(2r)} \,, \\ \|E(\boldsymbol{b})\|_F &\le e^{8s \log K - T^2/(2K)} \, \|\boldsymbol{b}\|_2 \sum_{i \in Q} \lambda_i^{(2s)} \,. \end{aligned} \tag{22}$$

(ii) For every $i, j \in Q$ and $1 \le r \le s$, the component weights satisfy

$$e^{-T^2} \le \mathbb{P}(T \le Y_i \le T+1) \le \lambda_i^{(2r)} \le e^{4r \log K - T^2/(3K)} \,, \tag{23}$$

$$K^{-(s-r)} \frac{\lambda_i^{(2s)}}{\lambda_j^{(2s)}} \le \frac{\lambda_i^{(2r)}}{\lambda_j^{(2r)}} \le K^{s-r} \frac{\lambda_i^{(2s)}}{\lambda_j^{(2s)}} \,. \tag{24}$$

(iii) The final lifted vectors are nearly orthogonal:

$$\left| \langle \boldsymbol{u}_i^{\otimes s}, \boldsymbol{u}_j^{\otimes s} \rangle \right| \le e^{-s/K} \qquad (i \ne j). \tag{25}$$

(iv) For every $i, j \in Q$, $\boldsymbol{a}_i$ is a positive multiple of $\boldsymbol{w}_i$, and

$$\frac{T}{2\sqrt{2s+1}} \frac{\|\boldsymbol{w}_i\|_2}{\sigma_i^2} \le \|\boldsymbol{a}_i\|_2 \le \frac{T+1}{\sqrt{2s+1}} \frac{\|\boldsymbol{w}_i\|_2}{\sigma_i^2} \,, \tag{26}$$

$$\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2^2 \ge \frac{1}{K} \max\left\{ \|\boldsymbol{a}_i\|_2^2, \|\boldsymbol{a}_j\|_2^2 \right\} \qquad (i \ne j) \,. \tag{27}$$

Theorem 5.1(i) approximately decomposes the Hermite moments as weighted rank-one tensor sums of the regressors with errors bounded relative to the total component weight. In Section 6, we estimate projections of these sums and use the rank-one matrices shared by the final even moment and the odd contractions to recover a block of normalized regressors.

Theorem 5.1(ii) bounds the component weights on the Gaussian tail scale and compares them with the slice probabilities. In Section 8, we use the slice-probability comparison to bound the peeling bias relative to the total component weight. In Section 6.2, the weight bounds set the estimation accuracy, and the comparison across degrees lets us bound the effect of the intermediate projections on the final even moment.

Theorem 5.1(iii) shows that the final lifted vectors are nearly orthogonal. In Section 6.3.1, we use this property to compare the eigenvalues of the even moment's rank-one sum with the component weights and choose a block on which to rescale the odd moments.

Theorem 5.1(iv) shows that each a_i is a positive multiple of w_i, with multiplier of order $T/(\sigma_i^2\sqrt{2s+1})$, and that distinct scaled regressors remain separated. In Section 6.3.2, we use their separation to recover the common eigenvectors from a random contraction. The lower bound on their norms controls the error when we normalize the estimated scaled regressors.

The rest of this section is dedicated to proving Theorem 5.1.

5.1 Decomposition of the Restricted Hermite Moments

To prove Theorem 5.1(i), we compare the restricted moment of the remaining maximum with the sum of the restricted moments of the individual responses. Section 5.1.1 computes each summand. Section 5.1.2 gives the probability and second-moment bounds needed to control the difference.

5.1.1 Single-Response Moments

For the contribution of component i, we first condition on $Y_i = t$ and then integrate over the slice $[T, T+1]$. The following identity gives the conditional Hermite moment.

Proposition 5.2 (Conditional Hermite Identity)**.** *Let $G \in \mathbb{R}^n$ and $Y \in \mathbb{R}$ be centered jointly Gaussian, with $G \sim N(0, I_n)$ and $\mathrm{Var}(Y) = \sigma^2 > 0$. Put $v = \mathbb{E}_{G,Y}[GY]$. For every integer $d \geq 0$ and $t \in \mathbb{R}$,*

$$\mathbb{E}_{G,Y}[\mathsf{H}_d(G) \mid Y = t] = \frac{\mathrm{He}_d(t/\sigma)}{\sigma^d\sqrt{d!}} v^{\otimes d}.$$

The proof in Appendix C.2 applies the scalar Gaussian noise identity to one-dimensional projections of G. In our model, $\mathbb{E}_{X,\xi_i}[XY_i] = w_i$. Integrating the conditional identity against $\phi_{\sigma_i^2}(t)$ over $[T, T+1]$ gives the following corollary.

Corollary 5.3 (Single-Response Hermite Identity)**.** *In the Gaussian self-selection model (Definition 1.1), for every $i \in [k]$ and integer $d \geq 0$,*

$$\mathbb{E}_{X,\xi_i}[\mathsf{H}_d(X)\mathbf{1}\{T \leq Y_i \leq T+1\}] = a_d(\sigma_i^2) {w_i}^{\otimes d}. \tag{28}$$

5.1.2 Moment Decomposition Error

The single-response sum counts Y_i whenever it lies in the slice, even if another remaining response is larger. We first bound the probability that another response exceeds t given $Y_i = t$, then bound the conditional second moment of the Hermite feature. Conditional on $Y_i = t$, the uncoveredness assumption ensures that the mean of each other response is below t by at least a constant fraction of t, while the boundedness assumption controls its conditional variance.

Lemma 5.4 (A Large Response Is Rarely Exceeded)**.** *Consider the Gaussian self-selection model (Definition 1.1) under the boundedness and uncoveredness assumptions (Assumption 1.2). Use the hyperparameters K, T, s from Hyperparameters 4.1. For every non-empty $Q \subseteq [k]$, $i \in Q$, and $t \in [T, T+1]$,*

$$\mathbb{P}(Y_{\max,Q} > t \mid Y_i = t) \leq k e^{-2t^2/K}.$$

Proof. By the choice of Hyperparameters 4.1,

$$K \geq \frac{4(1+B^2)^3}{(1+\Delta^2)^2}. \tag{29}$$

Conditional on $Y_i = t$, another response must exceed t for $Y_{\max,Q}$ to exceed t. Fix $j \in Q \setminus \{i\}$ and write $c_{ij} := \langle \boldsymbol{w}_i, \boldsymbol{w}_j \rangle$. Independence of the centered noises and $\boldsymbol{X}$ gives

$$\mathbb{E}_{\boldsymbol{X},\xi_i,\xi_j}[Y_i Y_j] = \boldsymbol{w}_i^\top \mathbb{E}_{\boldsymbol{X}}[\boldsymbol{X}\boldsymbol{X}^\top]\boldsymbol{w}_j = c_{ij}.$$

The conditional Gaussian formula (Proposition 4.3) gives

$$Y_j \mid Y_i = t \sim N\left(\frac{c_{ij}}{\sigma_i^2}t,\ \sigma_j^2 - \frac{c_{ij}^2}{\sigma_i^2}\right) =: N(\mu, \tau^2).$$

By the Gaussian tail bound (Proposition 4.2),

$$\begin{aligned}
t - \mu &= \frac{(\sigma_i^2 - c_{ij})t}{\sigma_i^2} \geq \frac{(1+\Delta^2)t}{\sigma_i^2} > 0\,, && \text{(uncoveredness (Assumption 1.2))}\\
\frac{(t-\mu)^2}{2\tau^2} &= \frac{(\sigma_i^2 - c_{ij})^2 t^2}{2\sigma_i^4(\sigma_j^2 - c_{ij}^2/\sigma_i^2)}\\
&\geq \frac{(\sigma_i^2 - c_{ij})^2 t^2}{2\sigma_i^4\sigma_j^2} && (\tau^2 \leq \sigma_j^2)\\
&\geq \frac{(1+\Delta^2)^2 t^2}{2(1+B^2)^3} && (\sigma_i^2, \sigma_j^2 \leq 1+B^2)\\
&\geq \frac{2t^2}{K}\,. && \text{(choice of } K \text{ (29))}
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{P}(Y_j > t \mid Y_i = t) &\leq \exp\left(-\frac{(t-\mu)^2}{2\tau^2}\right) \leq e^{-2t^2/K},\\
\mathbb{P}(Y_{\max,Q} > t \mid Y_i = t) &\leq \sum_{j \in Q\setminus\{i\}} \mathbb{P}(Y_j > t \mid Y_i = t) && \text{(union bound)}\\
&\leq (|Q|-1)e^{-2t^2/K} \leq k e^{-2t^2/K}. && \square
\end{aligned}$$

To bound the contribution of responses that enter the slice without attaining the maximum, we combine Lemma 5.4 with a conditional second-moment bound for the Hermite feature by Cauchy–Schwarz.

Proposition 5.5 (Conditional Hermite Second Moment). *Let $\boldsymbol{G} \in \mathbb{R}^n$ and $Y \in \mathbb{R}$ be centered jointly Gaussian, with $\boldsymbol{G} \sim N(\boldsymbol{0}, I_n)$ and $\mathrm{Var}(Y) = \sigma^2 > 0$. Put $\boldsymbol{v} = \mathbb{E}_{\boldsymbol{G},Y}[\boldsymbol{G}Y]$. For every integer $d \geq 0$, symmetric order-d tensor A, $t \in \mathbb{R}$, and $0 < \rho < 1/2$,*

$$\left(\mathbb{E}_{\boldsymbol{G},Y}[\langle \mathsf{H}_d(\boldsymbol{G}), \mathsf{A}\rangle^2 \mid Y = t]\right)^{1/2} \leq \frac{\rho^{-d/2}}{\sqrt{1-2\rho}} \exp\left(\frac{\rho\,\|\boldsymbol{v}\|_2^2\, t^2}{2\sigma^4(1-2\rho)}\right) \|\mathsf{A}\|_F\,.$$

We prove the proposition in Appendix C.3. Taking $(\boldsymbol{G}, Y) = (\boldsymbol{X}, Y_i)$ gives the following corollary.

Corollary 5.6 (Second Conditional Hermite Moment)**.** *Consider the Gaussian self-selection model (Definition 1.1) under the boundedness and uncoveredness assumptions (Assumption 1.2). Let $K \geq 2$, $T \geq 1$, and let $s \geq 1$ be an integer. For every $i \in [k]$, $1 \leq d \leq 2s+1$, $t \in [T, T+1]$, and symmetric order-d tensor A with $\|\mathsf{A}\|_F \leq 1$,*

$$\left(\mathbb{E}_{\boldsymbol{X},\xi_i}[\langle \mathsf{H}_d(\boldsymbol{X}), \mathsf{A}\rangle^2 \mid Y_i = t]\right)^{1/2} \leq e^{7s\log K + T^2/K^4}.$$

Proof. Apply Proposition 5.5 with $(\boldsymbol{G}, Y) = (\boldsymbol{X}, Y_i)$ and $\rho = K^{-4}$. Here $\boldsymbol{v} = \boldsymbol{w}_i$, and

$$\frac{\|\boldsymbol{w}_i\|_2}{\sigma_i^2} = \frac{\|\boldsymbol{w}_i\|_2}{1 + \|\boldsymbol{w}_i\|_2^2} \leq \frac{1}{2}.$$

Therefore

$$\begin{aligned}
\left(\mathbb{E}_{\boldsymbol{X},\xi_i}[\langle \mathsf{H}_d(\boldsymbol{X}), \mathsf{A}\rangle^2 \mid Y_i = t]\right)^{1/2} &\leq \frac{K^{2d}}{\sqrt{1-2K^{-4}}} \exp\left(\frac{t^2}{8K^4(1-2K^{-4})}\right) && (\|\mathsf{A}\|_F \leq 1)\\
&\leq \sqrt{2}\, K^{2d} e^{t^2/(4K^4)} && (K^{-4} \leq 1/16)\\
&\leq e^{7s\log K + T^2/K^4}. && (d \leq 3s,\ \sqrt{2} \leq K^s,\ t \leq 2T)
\end{aligned}$$

□

5.2 Bounds on the Component Weights

To prove Theorem 5.1(ii), we first establish positivity and ratio bounds for the restricted moment coefficients. We use these bounds to compare each component weight with its slice probability, then to compare the relative weights of two components across degrees.

5.2.1 Restricted Moment Coefficients

The weights depend on the even coefficients, while the scaled regressors depend on the ratio of two consecutive coefficients. We first bound these ratios for arbitrary degrees.

Lemma 5.7 (Restricted Moment Coefficient Estimates)**.** *Use the choice of K from Hyperparameters 4.1. Let $s \geq 1$ be an integer and suppose $T > 0$ and $T^2 \geq Ks$. For every $\sigma^2 \in [1+\Delta^2, 1+B^2]$, the coefficients $a_d(\sigma^2)$ are positive for $0 \leq d \leq 2s+2$. Moreover, for every pair of integers $0 \leq d_1 \leq d_2 \leq 2s+2$,*

$$\sqrt{\frac{d_1!}{d_2!}}\left(\frac{T}{2\sigma^2}\right)^{d_2-d_1} \leq \frac{a_{d_2}(\sigma^2)}{a_{d_1}(\sigma^2)} \leq \sqrt{\frac{d_1!}{d_2!}}\left(\frac{T+1}{\sigma^2}\right)^{d_2-d_1}. \tag{30}$$

Proof. The requirement on K gives $K \geq 16(1+B^2)$. Consequently,

$$T^2 \geq Ks \geq 16s\sigma^2 \geq 4(2s+2)\sigma^2.$$

Fix $x \geq 2\sqrt{2s+2}$. We use the standard Hermite recurrence [DLMF, Table 18.9.1],

$$\mathrm{He}_{j+1}(x) = x\,\mathrm{He}_j(x) - j\,\mathrm{He}_{j-1}(x).$$

Starting from $\mathrm{He}_0(x) = 1$ and $\mathrm{He}_1(x) = x$, we show by induction that the ratios $q_j = \mathrm{He}_j(x)/\mathrm{He}_{j-1}(x)$ are well-defined and lie in $[x/2, x]$ for $1 \le j \le 2s+2$. Indeed, $q_1 = x$, and for $1 \le j < 2s+2$ the recurrence gives

$$x/2 \le x - \frac{2j}{x} \le q_{j+1} = x - \frac{j}{q_j} \le x.$$

Multiplying these ratios gives, for $1 \le d \le 2s+2$,

$$(x/2)^d \le \mathrm{He}_d(x) = \prod_{j=1}^{d} q_j \le x^d.$$

The threshold bound makes these estimates applicable throughout the slice, so $\mathrm{He}_d(t/\sigma) > 0$ for $1 \le d \le 2s+2$. The same holds for $d = 0$, since $\mathrm{He}_0 = 1$. Thus $a_d(\sigma^2) > 0$ for $0 \le d \le 2s+2$.

We next compare coefficients of different degrees. Fix $0 \le d_1 \le d_2 \le 2s+2$. For $t \in [T, T+1]$, the recurrence bounds give

$$\frac{\mathrm{He}_{d_2}(t/\sigma)}{\mathrm{He}_{d_1}(t/\sigma)} = \prod_{j=d_1+1}^{d_2} \frac{\mathrm{He}_j(t/\sigma)}{\mathrm{He}_{j-1}(t/\sigma)} \in \left[\left(\frac{t}{2\sigma}\right)^{d_2-d_1}, \left(\frac{t}{\sigma}\right)^{d_2-d_1} \right].$$

When $d_1 = d_2$, the product is empty and equals one. Including the normalization factors and using $T \le t \le T+1$ therefore gives

$$\sqrt{\frac{d_1!}{d_2!}} \left(\frac{T}{2\sigma^2}\right)^{d_2-d_1} \le \frac{\sigma^{-d_2}\,\mathrm{He}_{d_2}(t/\sigma)/\sqrt{d_2!}}{\sigma^{-d_1}\,\mathrm{He}_{d_1}(t/\sigma)/\sqrt{d_1!}} \le \sqrt{\frac{d_1!}{d_2!}} \left(\frac{T+1}{\sigma^2}\right)^{d_2-d_1}.$$

Multiply by the positive function $\phi_{\sigma^2}(t)\sigma^{-d_1}\,\mathrm{He}_{d_1}(t/\sigma)/\sqrt{d_1!}$ and integrate over $[T, T+1]$. This yields

$$\sqrt{\frac{d_1!}{d_2!}} \left(\frac{T}{2\sigma^2}\right)^{d_2-d_1} a_{d_1}(\sigma^2) \le a_{d_2}(\sigma^2) \le \sqrt{\frac{d_1!}{d_2!}} \left(\frac{T+1}{\sigma^2}\right)^{d_2-d_1} a_{d_1}(\sigma^2).$$

Dividing by $a_{d_1}(\sigma^2) > 0$ proves (30). □

5.2.2 Slice Probabilities and Weights

The degree-zero coefficient is the slice probability. Comparing it with the degree-$2r$ coefficient gives the weight bounds in Theorem 5.1(ii) and the comparison needed for the error bound in Theorem 5.1(i).

Lemma 5.8 (Even-Moment Weight Bounds)**.** *Consider the Gaussian self-selection model (Definition 1.1) under the boundedness and uncoveredness assumptions (Assumption 1.2). Use the choice of K from Hyperparameters 4.1. Let $s \ge 1$ be an integer and suppose $T > 0$ and $T^2 \ge Ks$. For every $i \in [k]$ and $1 \le r \le s$,*

$$\mathbb{P}(T \le Y_i \le T+1) \le \lambda_i^{(2r)} \le e^{4r\log K + T^2/(2K^4)}\,\mathbb{P}(T \le Y_i \le T+1). \tag{31}$$

Consequently,

$$e^{-T^2} \le \lambda_i^{(2r)} \le e^{4r\log K - T^2/(3K)}. \tag{32}$$

Proof. Fix $i \in [k]$ and $1 \le r \le s$. We first compare the weight with the slice probability. Since $\mathrm{He}_0 = 1$,

$$a_0(\sigma_i^2) = \mathbb{P}(T \le Y_i \le T+1) > 0.$$

Consequently,

$$\begin{aligned}
\frac{\lambda_i^{(2r)}}{a_0(\sigma_i^2)} &= \frac{a_{2r}(\sigma_i^2)}{a_0(\sigma_i^2)} \|\boldsymbol{w}_i\|_2^{2r} \\
&\ge \frac{(T \|\boldsymbol{w}_i\|_2 / (2\sigma_i^2))^{2r}}{\sqrt{(2r)!}} && \text{(by Lemma 5.7)} \\
&\ge \left(\frac{T^2\Delta^2}{8r(1+B^2)^2}\right)^r && ((2r)! \le (2r)^{2r};\ \|\boldsymbol{w}_i\|_2 \ge \Delta;\ \sigma_i^2 \le 1+B^2) \\
&\ge \left(\frac{2T^2}{Kr}\right)^r && \text{(choice of } K\text{: } K\Delta^2 \ge 16(1+B^2)^2) \\
&\ge 1. && (T^2 \ge Ks,\ r \le s)
\end{aligned}$$

Multiplying by the positive coefficient $a_0(\sigma_i^2)$ proves the lower comparison in (31). For the upper comparison, use $2\|\boldsymbol{w}_i\|_2 \le 1 + \|\boldsymbol{w}_i\|_2^2 = \sigma_i^2$ and $T \ge 1$. Then

$$\begin{aligned}
\frac{\lambda_i^{(2r)}}{a_0(\sigma_i^2)} &\le \frac{\left((T+1)\|\boldsymbol{w}_i\|_2 / \sigma_i^2\right)^{2r}}{\sqrt{(2r)!}} && \text{(by Lemma 5.7)} \\
&\le \frac{T^{2r}}{\sqrt{(2r)!}} && (2\|\boldsymbol{w}_i\|_2 \le \sigma_i^2,\ T+1 \le 2T) \\
&= K^{4r}\left(\frac{(T^2/K^4)^{2r}}{(2r)!}\right)^{1/2} \\
&\le K^{4r}\left(\sum_{\ell=0}^{\infty}\frac{(T^2/K^4)^\ell}{\ell!}\right)^{1/2} && \text{(all terms are nonnegative)} \\
&= e^{4r\log K + T^2/(2K^4)}.
\end{aligned}$$

Multiplying by $a_0(\sigma_i^2) > 0$ proves the upper comparison in (31).

It remains to prove (32). The hypotheses give $K \ge 128$ and $T^2 \ge K$, so $T \ge 5$. For the lower bound,

$$\begin{aligned}
\mathbb{P}(T \le Y_i \le T+1) &\ge \phi_{\sigma_i^2}(T+1) && \text{(the density decreases on } [T, T+1]) \\
&\ge \exp\left(-\frac{(T+1)^2}{2} - \frac{1}{2}\log(2\pi K)\right) && (1 \le \sigma_i^2 \le K) \\
&\ge e^{-3T^2/4 - K/4} && (T \ge 5,\ K \ge 128) \\
&\ge e^{-T^2}. && (T^2 \ge K)
\end{aligned}$$

Combining this with (31) gives the lower bound on $\lambda_i^{(2r)}$. For the upper bound,

$$\begin{aligned}
\lambda_i^{(2r)} &\le e^{4r\log K + T^2/(2K^4)}\phi_{\sigma_i^2}(T) && \text{(by (31); decreasing density)} \\
&\le e^{4r\log K + T^2/(2K^4) - T^2/(2K)} && (1 \le \sigma_i^2 \le K) \\
&\le e^{4r\log K - T^2/(3K)}. && (K^3 \ge 3)
\end{aligned}$$

Together, these bounds prove the claim. □

5.2.3 Weights Across Degrees

We have bounded each weight at a fixed degree. For the intermediate projections, we also need to know how the ratio of two component weights changes between levels r and s.

Lemma 5.9 (Weights Across Degrees)**.** *Consider the Gaussian self-selection model (Definition 1.1) under the boundedness and uncoveredness assumptions (Assumption 1.2). Use the hyperparameters K, T, s from Hyperparameters 4.1. For every $i, j \in [k]$ and $1 \le r \le s$,*

$$\left| \log \frac{\lambda_j^{(2r)} / \lambda_i^{(2r)}}{\lambda_j^{(2s)} / \lambda_i^{(2s)}} \right| \le (s - r) \log K.$$

Proof. By the choice of Hyperparameters 4.1,

$$K \ge \frac{16(1 + B^2)^3}{\Delta^4}. \tag{33}$$

Fix $i, j \in [k]$ and $1 \le r \le s$. The coefficients are positive by Lemma 5.7. We compare the two components directly:

$$\begin{aligned}
\frac{\lambda_j^{(2r)} / \lambda_i^{(2r)}}{\lambda_j^{(2s)} / \lambda_i^{(2s)}} &= \frac{a_{2s}(\sigma_i^2) / a_{2r}(\sigma_i^2)}{a_{2s}(\sigma_j^2) / a_{2r}(\sigma_j^2)} \left(\frac{\|\boldsymbol{w}_i\|_2}{\|\boldsymbol{w}_j\|_2} \right)^{2(s-r)} \\
&\le \frac{\sqrt{\frac{(2r)!}{(2s)!}} \left((T+1) \|\boldsymbol{w}_i\|_2 / \sigma_i^2 \right)^{2(s-r)}}{\sqrt{\frac{(2r)!}{(2s)!}} \left(T \|\boldsymbol{w}_j\|_2 / (2\sigma_j^2) \right)^{2(s-r)}} && \text{(by Lemma 5.7)} \\
&= \left(\frac{2(T+1)}{T} \frac{\|\boldsymbol{w}_i\|_2}{\|\boldsymbol{w}_j\|_2} \frac{\sigma_j^2}{\sigma_i^2} \right)^{2(s-r)} \\
&\le \left(\frac{3B(1+B^2)}{\Delta} \right)^{2(s-r)} && \text{(by Assumption 1.2; } T \ge 2) \\
&\le \left(\frac{16(1+B^2)^3}{\Delta^4} \right)^{s-r} && (9B^2\Delta^2 \le 16(1+B^2)) \\
&\le K^{s-r}. && \text{(choice of } K \text{ (33))} \quad (34)
\end{aligned}$$

Interchanging i and j gives the same bound for the reciprocal. Taking logarithms proves the claim. □

5.3 Proof of Theorem 5.1

We first combine the single-response identity and the error estimates from Sections 5.1.1 and 5.1.2 to obtain the even and odd decompositions. The weight comparisons give Theorem 5.1(ii). We finish by applying uncoveredness to the lifted vectors and the coefficient ratio bounds to the scaled regressors.

Proof of Theorem 5.1. We first prove Theorem 5.1(i) by identifying the rank-one terms and bounding the contribution of latent responses that enter the slice without attaining the maximum.

Let I_Q be the maximizing index within Q, using the global tie rule. Decomposing the slice by this index gives

$$\mathbf{1}\{T \le Y_{\max,Q} \le T+1\} = \sum_{i \in Q} \mathbf{1}\{I_Q = i,\ T \le Y_i \le T+1\}.$$

For $1 \le d \le 2s+1$, define

$$\mathsf{R}_d^Q := \sum_{i \in Q} \mathbb{E}_{\boldsymbol{X},\xi_Q}[\mathsf{H}_d(\boldsymbol{X})\mathbf{1}\{T \le Y_i \le T+1,\ I_Q \ne i\}].$$

Since $\mathbf{1}\{I_Q = i\} = 1 - \mathbf{1}\{I_Q \ne i\}$, we obtain

$$\begin{aligned}
\mathsf{T}_d^Q &= \sum_{i \in Q} \mathbb{E}_{\boldsymbol{X},\xi_Q}[\mathsf{H}_d(\boldsymbol{X})\mathbf{1}\{T \le Y_i \le T+1\}\mathbf{1}\{I_Q = i\}] \\
&= \sum_{i \in Q} \mathbb{E}_{\boldsymbol{X},\xi_i}[\mathsf{H}_d(\boldsymbol{X})\mathbf{1}\{T \le Y_i \le T+1\}] - \mathsf{R}_d^Q \\
&= \sum_{i \in Q} a_d(\sigma_i^2)\boldsymbol{w}_i^{\otimes d} - \mathsf{R}_d^Q. \qquad (35)
\end{aligned}$$

The last equality applies Corollary 5.3 to each latent response. Thus R_d^Q collects the contributions of responses that enter the slice without attaining the maximum.

We now bound the Frobenius norm of the error tensor. Fix a symmetric order-d tensor A with $\|\mathsf{A}\|_F \le 1$. Conditional on $Y_i = t$, ties have probability zero, so $\mathbb{P}(I_Q \ne i \mid Y_i = t) = \mathbb{P}(Y_{\max,Q} > t \mid Y_i = t)$. For $t \in [T, T+1]$, Lemma 5.4 bounds the probability that another response exceeds t by $ke^{-2t^2/K}$ conditional on $Y_i = t$. Under the same conditioning, Corollary 5.6 bounds the L^2 norm of the Hermite pairing by $e^{7s\log K + T^2/K^4}$. Conditioning on each latent response gives

$$\begin{aligned}
\left|\left\langle \mathsf{R}_d^Q, \mathsf{A}\right\rangle\right| &= \left|\sum_{i \in Q} \int_T^{T+1} \mathbb{E}_{\boldsymbol{X},\xi_Q}[\langle \mathsf{H}_d(\boldsymbol{X}), \mathsf{A}\rangle\, \mathbf{1}\{I_Q \ne i\} \mid Y_i = t]\, \phi_{\sigma_i^2}(t)\, \mathrm{d}t\right| \\
&\le \sum_{i \in Q} \int_T^{T+1} \mathbb{E}_{\boldsymbol{X},\xi_Q}[|\langle \mathsf{H}_d(\boldsymbol{X}), \mathsf{A}\rangle|\, \mathbf{1}\{I_Q \ne i\} \mid Y_i = t]\, \phi_{\sigma_i^2}(t)\, \mathrm{d}t \qquad \text{(triangle inequality)} \\
&\le \sum_{i \in Q} \int_T^{T+1} \left(\mathbb{E}_{\boldsymbol{X},\xi_i}[\langle \mathsf{H}_d(\boldsymbol{X}), \mathsf{A}\rangle^2 \mid Y_i = t]\right)^{1/2} \mathbb{P}(I_Q \ne i \mid Y_i = t)^{1/2} \phi_{\sigma_i^2}(t)\, \mathrm{d}t \\
&\qquad \text{(conditional Cauchy–Schwarz)} \\
&\le e^{7s\log K + \frac{1}{2}\log k + T^2/K^4} \sum_{i \in Q} \int_T^{T+1} e^{-t^2/K} \phi_{\sigma_i^2}(t)\, \mathrm{d}t \qquad \text{(by Corollary 5.6 and Lemma 5.4)} \\
&\le e^{8s\log K + T^2/K^4} \sum_{i \in Q} \int_T^{T+1} e^{-t^2/K} \phi_{\sigma_i^2}(t)\, \mathrm{d}t \qquad \text{(choice of } s\text{: } \log k \le s;\ \log K \ge 1/2) \\
&\le e^{8s\log K} \sum_{i \in Q} \int_T^{T+1} e^{-t^2/(2K)} \phi_{\sigma_i^2}(t)\, \mathrm{d}t \qquad (t \ge T,\ K^3 \ge 2) \\
&\le e^{8s\log K - T^2/(2K)} \sum_{i \in Q} \mathbb{P}(T \le Y_i \le T+1). \qquad (t \ge T)
\end{aligned}$$

Since R_d^Q is symmetric, taking the supremum over symmetric A with $\|\mathsf{A}\|_F \le 1$ bounds its Frobenius norm. By Lemma 5.8, each slice probability is at most $\lambda_i^{(2r)}$ for every $1 \le r \le s$. Hence

$$\left\|\mathsf{R}_d^Q\right\|_F \le e^{8s\log K - T^2/(2K)} \sum_{i\in Q} \lambda_i^{(2r)} \qquad (1 \le d \le 2s+1,\ 1 \le r \le s). \tag{36}$$

We now apply flattening to obtain the even moment matrices. For $1 \le r \le s$, recall that $\boldsymbol{w}_i = \|\boldsymbol{w}_i\|_2 \boldsymbol{u}_i$, which gives

$$\mathrm{Mat}_r(\boldsymbol{w}_i^{\otimes 2r}) = \|\boldsymbol{w}_i\|_2^{2r}\, \boldsymbol{u}_i^{\otimes r}(\boldsymbol{u}_i^{\otimes r})^\top.$$

Applying Mat_r to the winner decomposition (35) with $d = 2r$ therefore gives

$$\begin{aligned}\mathrm{Mat}_r(\mathsf{T}_{2r}^Q) &= \sum_{i\in Q} a_{2r}(\sigma_i^2)\, \|\boldsymbol{w}_i\|_2^{2r}\, \boldsymbol{u}_i^{\otimes r}(\boldsymbol{u}_i^{\otimes r})^\top - \mathrm{Mat}_r(\mathsf{R}_{2r}^Q)\\ &= \sum_{i\in Q} \lambda_i^{(2r)} \boldsymbol{u}_i^{\otimes r}(\boldsymbol{u}_i^{\otimes r})^\top + E_r.\end{aligned}$$

Here $E_r = -\mathrm{Mat}_r(\mathsf{R}_{2r}^Q)$. The last equality restates the weight definition in (18). Flattening preserves the Frobenius norm, so (36) with $d = 2r$ gives the even error bound in Theorem 5.1(i).

For the odd moment, fix $\boldsymbol{b} \in \mathbb{R}^n$. Contracting one index gives

$$\boldsymbol{w}_i^{\otimes(2s+1)} \lrcorner\, \boldsymbol{b} = \langle \boldsymbol{b}, \boldsymbol{w}_i\rangle\, \boldsymbol{w}_i^{\otimes 2s}.$$

Contracting one index against $\boldsymbol{b}$ and then applying Mat_s to the winner decomposition (35) with $d = 2s+1$ gives

$$\begin{aligned}&\mathrm{Mat}_s(\mathsf{T}_{2s+1}^Q \lrcorner\, \boldsymbol{b})\\ &\quad= \sum_{i\in Q} a_{2s+1}(\sigma_i^2)\, \langle \boldsymbol{b}, \boldsymbol{w}_i\rangle\, \|\boldsymbol{w}_i\|_2^{2s}\, \boldsymbol{u}_i^{\otimes s}(\boldsymbol{u}_i^{\otimes s})^\top - \mathrm{Mat}_s(\mathsf{R}_{2s+1}^Q \lrcorner\, \boldsymbol{b})\\ &\quad= \sum_{i\in Q} \lambda_i^{(2s)}\, \langle \boldsymbol{b}, \boldsymbol{a}_i\rangle\, \boldsymbol{u}_i^{\otimes s}(\boldsymbol{u}_i^{\otimes s})^\top + E(\boldsymbol{b}).\end{aligned}$$

Here $E(\boldsymbol{b}) = -\mathrm{Mat}_s(\mathsf{R}_{2s+1}^Q \lrcorner\, \boldsymbol{b})$. The last equality restates the definitions in (18) and (19), since

$$a_{2s+1}(\sigma_i^2)\, \langle \boldsymbol{b}, \boldsymbol{w}_i\rangle\, \|\boldsymbol{w}_i\|_2^{2s} = a_{2s}(\sigma_i^2)\, \|\boldsymbol{w}_i\|_2^{2s} \left\langle \boldsymbol{b}, \frac{a_{2s+1}(\sigma_i^2)}{a_{2s}(\sigma_i^2)} \boldsymbol{w}_i \right\rangle = \lambda_i^{(2s)}\, \langle \boldsymbol{b}, \boldsymbol{a}_i\rangle\,.$$

By Cauchy–Schwarz in the contracted index,

$$\|E(\boldsymbol{b})\|_F = \left\|\mathsf{R}_{2s+1}^Q \lrcorner\, \boldsymbol{b}\right\|_F \le \left\|\mathsf{R}_{2s+1}^Q\right\|_F \|\boldsymbol{b}\|_2\,.$$

Taking $d = 2s+1$ and $r = s$ in (36) gives the odd error bound and proves Theorem 5.1(i).

Proof of Theorem 5.1(ii). The proof of Lemma 5.8 bounds the slice probability below by e^{-T^2} and above by $\lambda_i^{(2r)}$, and bounds the weight above by $e^{4r\log K - T^2/(3K)}$. By Lemma 5.9, the log ratio of two component weights changes by at most $(s-r)\log K$ between levels r and s. Exponentiating gives the ratio bounds and proves Theorem 5.1(ii).

Proof of Theorem 5.1(iii). By Assumption 1.2, for distinct i, j,

$$\left|\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle\right| \le 1 - \frac{\Delta^2}{B^2}.$$

Taking tensor powers gives

$$\left|\langle \boldsymbol{u}_i^{\otimes s}, \boldsymbol{u}_j^{\otimes s} \rangle\right| = \left|\langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle\right|^s \le \left(1 - \frac{\Delta^2}{B^2}\right)^s \le e^{-s\Delta^2/B^2} \le e^{-s/K}.$$

The last inequality uses $K\Delta^2/B^2 \ge 1$, a consequence of the choice of Hyperparameters 4.1. This proves Theorem 5.1(iii).

Proof of Theorem 5.1(iv). By Lemma 5.7,

$$0 < \frac{T}{2\sigma_i^2\sqrt{2s+1}} \le \frac{a_{2s+1}(\sigma_i^2)}{a_{2s}(\sigma_i^2)} \le \frac{T+1}{\sigma_i^2\sqrt{2s+1}}.$$

Multiplying by $\|\boldsymbol{w}_i\|_2$ gives the norm bounds in Theorem 5.1(iv). In particular, $\boldsymbol{a}_i / \|\boldsymbol{a}_i\|_2 = \boldsymbol{u}_i$.

To prove separation, project $\boldsymbol{a}_i - \boldsymbol{a}_j$ onto $\boldsymbol{u}_j^\perp$. Since $\boldsymbol{a}_j$ lies in the direction $\boldsymbol{u}_j$,

$$\begin{aligned}
\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2^2 &\ge \left\|(I_n - \boldsymbol{u}_j\boldsymbol{u}_j^\top)(\boldsymbol{a}_i - \boldsymbol{a}_j)\right\|_2^2 && \text{(orthogonal projection)} \\
&= \|\boldsymbol{a}_i\|_2^2 \left(1 - \langle \boldsymbol{u}_i, \boldsymbol{u}_j \rangle^2\right) \\
&\ge \|\boldsymbol{a}_i\|_2^2 \left(1 - (1 - \Delta^2/B^2)^2\right) && \text{(by Assumption 1.2)} \\
&\ge \frac{\Delta^2}{B^2} \|\boldsymbol{a}_i\|_2^2 && (0 < \Delta^2/B^2 \le 1) \\
&\ge \frac{1}{K} \|\boldsymbol{a}_i\|_2^2. && \text{(choice of } K \text{ (Hyperparameters 4.1): } K\Delta^2 \ge B^2)
\end{aligned}$$

Interchanging i and j gives the separation bound in Theorem 5.1(iv) and completes the proof. □

6 Implicit Tensor Decomposition

Section 5 shows that the restricted Hermite moments of the maximum response among the unrecovered regressors are close to weighted sums of tensor powers of those regressors. At degree $s = \Theta(\log n)$, the lifted vectors $\boldsymbol{u}_i^{\otimes s}$ are nearly orthogonal. In this section, we present a tensor decomposition algorithm for these implicit high-order tensors. More specifically, given a set $\widehat{\mathcal{R}} \subseteq \mathbb{R}^n$ of recovered regressors, with a bijection match : $\widehat{\mathcal{R}} \to R \subsetneq [k]$, we will recover unit vectors parallel to some subset of the remaining regressors $\boldsymbol{w}_i, i \in Q := [k] \setminus R$.

For ImplicitTensorDecomposition, it suffices that the recovered regressors have error at most $e^{-K^6 s}$.

Theorem 6.1 (Implicit Tensor Decomposition)**.** *Consider the Gaussian self-selection model (Definition 1.1) under the boundedness and uncoveredness assumptions (Assumption 1.2). Suppose the recovered list $\widehat{\mathcal{R}}$ is matched bijectively to $R \subsetneq [k]$ and*

$$\left\|\widehat{\boldsymbol{w}} - \boldsymbol{w}_{\mathrm{match}(\widehat{\boldsymbol{w}})}\right\|_2 \le e^{-K^6 s} \qquad (\widehat{\boldsymbol{w}} \in \widehat{\mathcal{R}}).$$

Put $Q = [k] \setminus R$. Given fresh samples independent of $\widehat{\mathcal{R}}$ and a confidence parameter $\delta \in (0, 1/3)$, ImplicitTensorDecomposition *(Algorithm 6.4) returns unit vectors $\widehat{\boldsymbol{u}}_1, \dots, \widehat{\boldsymbol{u}}_{i_\star}$, where $1 \le i_\star \le |Q|$, such that, with probability at least $1 - \delta$, there is an injection $\pi : [i_\star] \to Q$ satisfying*

$$\max_{\ell \in [i_\star]} \left\| \widehat{\boldsymbol{u}}_\ell - \frac{\boldsymbol{w}_{\pi(\ell)}}{\left\| \boldsymbol{w}_{\pi(\ell)} \right\|_2} \right\|_2 \le e^{-s/(2K)} \le \frac{\Delta^2}{16B^2}. \tag{37}$$

Its sample complexity is

$$\widetilde{O}\Big((nk)^{O(K^{10})} \log(1/\delta) \Big),$$

and its running time is

$$(nk)^{\mathrm{poly}(K)} \log(1/\delta).$$

The rest of this section is dedicated to proving Theorem 6.1.

6.1 Proof Roadmap

Our implicit tensor decomposition algorithm consists of three key subroutines. The first algorithm is an iterative projection algorithm that produces projections of high-order Hermite moments without having to explicitly form the tensors. Given the projected moments, we construct a family of matrices that encode the regressors within its eigenvalues. Finally, we decode the regressors using an eigendecomposition.

Iterative Projections (Section 6.2). Given the recovered list $\widehat{\mathcal{R}}$, fresh samples, a sampling accuracy, and a failure probability, EstimateProjectedMoments returns a projected even-moment estimate $\widehat{A}$ and projected odd moment estimates $\widehat{M}_1, \dots, \widehat{M}_n$. We quantify what it means to estimate below.

For the remaining set Q, write

$$\begin{aligned} A_Q^{(r)} &:= \sum_{i \in Q} \lambda_i^{(2r)} \boldsymbol{u}_i^{\otimes r} (\boldsymbol{u}_i^{\otimes r})^\top, \qquad 1 \le r \le s, \\ M_Q^{(s)}(\boldsymbol{b}) &:= \sum_{i \in Q} \lambda_i^{(2s)} \langle \boldsymbol{b}, \boldsymbol{a}_i \rangle \, \boldsymbol{u}_i^{\otimes s} (\boldsymbol{u}_i^{\otimes s})^\top, \qquad \boldsymbol{b} \in \mathbb{R}^n. \end{aligned} \tag{38}$$

Definition 6.2 (Projected Moment Estimate)**.** Fix a non-empty $Q \subseteq [k]$ and the target moment matrices $A_Q^{(s)}$ and $M_Q^{(s)}(\cdot)$ from (38). Let $C_s : (\mathbb{R}^n)^{\otimes s} \to \mathbb{R}^{p_s}$ have orthonormal rows. For $\varepsilon \ge 0$, we call symmetric matrices $\widehat{A}, \widehat{M}_1, \dots, \widehat{M}_n \in \mathbb{R}^{p_s \times p_s}$ an *ε-estimate of the projected moment matrices for Q relative to C_s* if the following bounds hold:

(i) The even estimate satisfies

$$\left\| \widehat{A} - C_s A_Q^{(s)} C_s^\top \right\|_F \le \varepsilon \, \mathrm{tr}(A_Q^{(s)}). \tag{39}$$

(ii) For every $\boldsymbol{b} \in \mathbb{R}^n$, the odd estimates satisfy

$$\left\| \sum_{\ell=1}^n b_\ell \widehat{M}_\ell - C_s M_Q^{(s)}(\boldsymbol{b}) C_s^\top \right\|_F \le \varepsilon \, \mathrm{tr}(A_Q^{(s)}) \, \|\boldsymbol{b}\|_2 . \tag{40}$$

We are now ready to state the guarantee of the first subroutine.

Theorem 6.3 (Projected-Moment Accuracy)**.** *Consider the Gaussian self-selection model (Definition 1.1) under the boundedness and uncoveredness assumptions (Assumption 1.2), with $k \geq 2$ and the hyperparameters from Hyperparameters 4.1. Fix a recovered list $\widehat{\mathcal{R}}$ matched bijectively to $R \subsetneq [k]$ such that*

$$\left\| \widehat{\boldsymbol{w}} - \boldsymbol{w}_{\mathrm{match}(\widehat{\boldsymbol{w}})} \right\|_2 \leq e^{-K^6 s} \qquad (\widehat{\boldsymbol{w}} \in \widehat{\mathcal{R}}),$$

and put $Q = [k] \setminus R$.

For $\delta \in (0, 1/3)$, run EstimateProjectedMoments *(Algorithm 6.1) with sampling accuracy $e^{-K^6 s}$, failure probability δ, and fresh observations independent of $\widehat{\mathcal{R}}$. With probability at least $1 - \delta$, the returned matrices $\widehat{A}, \widehat{M}_1, \dots, \widehat{M}_n$ form an $e^{-K^2 s}$-estimate of the projected moment matrices for Q relative to the constructed map C_s (Definition 6.2), and*

$$\left\| (I - C_s^\top C_s)(A_Q^{(s)})^{1/2} \right\|_{\mathrm{op}}^2 \leq 2K^{s-1} e^{-K^2 s} \operatorname{tr}(A_Q^{(s)}).$$

Its sample complexity is

$$\widetilde{O}\Big((nk)^{O(K^{10})} \log(1/\delta) \Big),$$

and its running time is

$$(nk)^{\mathrm{poly}(K)} \log(1/\delta).$$

In order for Theorem 6.3 to compute these projected moments in polynomial time, it must avoid explicitly materializing high-order tensors. This is accomplished by iteratively building the projections degree by degree.

Observable-Operator Construction (Section 6.3.1). Given $\widehat{A}$ and $\widehat{M}_1, \dots, \widehat{M}_n$, ConstructObservableOperators returns a family of matrices $\widehat{D}_1, \dots, \widehat{D}_n$ that encode a subset of the unrecovered regressors. We formalize the precise spectral condition below.

Definition 6.4 (Approximate Observable-Operator Family)**.** Fix a non-empty set J and vectors $(\boldsymbol{a}_i)_{i \in J}$ in $\mathbb{R}^n$. For $\varepsilon_{\mathrm{fam}} \geq 0$, we call symmetric matrices $\widehat{D}_1, \dots, \widehat{D}_n \in \mathbb{R}^{|J| \times |J|}$ an $\varepsilon_{\mathrm{fam}}$-approximate observable-operator family for $(\boldsymbol{a}_i)_{i \in J}$ if there is an orthonormal basis $(\boldsymbol{z}_i)_{i \in J}$ of $\mathbb{R}^{|J|}$ such that, for every $\boldsymbol{b} \in \mathbb{R}^n$,

$$\left\| \sum_{\ell=1}^{n} b_\ell \widehat{D}_\ell - \sum_{i \in J} \langle \boldsymbol{b}, \boldsymbol{a}_i \rangle \, \boldsymbol{z}_i \boldsymbol{z}_i^\top \right\|_F \leq \varepsilon_{\mathrm{fam}} \|\boldsymbol{b}\|_2 . \tag{41}$$

We now state the guarantees of the second subroutine.

Theorem 6.5 (Observable-Operator Construction)**.** *Fix a non-empty $Q \subseteq [k]$. Consider the rank-one sums $A_Q^{(s)}, M_Q^{(s)}(\boldsymbol{b})$ from (38). Let $C_s \in \mathbb{R}^{p_s \times n^s}$ have orthonormal rows for some $1 \leq p_s \leq |Q|$ and $\varepsilon_{\mathrm{mom}}, \varepsilon_{\mathrm{proj}}, \varepsilon_\perp \geq 0$. Suppose the following conditions hold.*

(a) For every $i \in Q$, $\|\boldsymbol{a}_i\|_2 \leq K^3$.

(b) The lifted vectors satisfy

$$\left\| \left(\langle \boldsymbol{u}_i^{\otimes s}, \boldsymbol{u}_j^{\otimes s} \rangle \right)_{i,j \in Q} - I \right\|_{\mathrm{op}} \leq \varepsilon_\perp. \tag{42}$$

(c) *The matrices* $\widehat{A}, \widehat{M}_1, \dots, \widehat{M}_n$ *form an* $\varepsilon_{\mathrm{mom}}$*-estimate of the projected moment matrices for* Q *relative to* C_s *(Definition 6.2).*

(d) *The projection error satisfies*

$$\left\| (I - C_s^\top C_s)(A_Q^{(s)})^{1/2} \right\|_{\mathrm{op}}^2 \le \varepsilon_{\mathrm{proj}} \operatorname{tr}(A_Q^{(s)}). \tag{43}$$

(e) *The errors satisfy* $\varepsilon_{\mathrm{mom}} + \varepsilon_{\mathrm{proj}} + \varepsilon_\perp \le \frac{1}{16k^2}$.

Then ConstructObservableOperators *(Algorithm 6.2), given these matrices, returns a* $16K^3k^3(\varepsilon_{\mathrm{mom}} + \varepsilon_{\mathrm{proj}} + \varepsilon_\perp)$*-approximate observable-operator family (Definition 6.4) for* $(\boldsymbol{a}_i)_{i\in J}$ *indexed by some* $J \subseteq Q$*. Moreover, the routine runs in* $\mathrm{poly}(n,k)$ *time.*

Theorem 6.5 is the main bottleneck which prevents our algorithm from recovering all regressors at once. This is because the construction roughly requires the eigenvalues of $\widehat{A}$ to be on the same relative scale. However, due to the irreducible bias, we have no control over the smallest eigenvalues and must truncate to the top eigenspace instead.

Recovering Normalized Regressors (Section 6.3.2). Given an approximate observable-operator family $\widehat{D}_1, \dots, \widehat{D}_n$ of dimension $i_\star \ge 2$ and a failure probability, RecoverDirections returns $i_\star$ unit vectors estimating the normalized regressors in the block up to permutation.

Theorem 6.6 (Normalized-Regressor Recovery)**.** *Let* $J \subseteq [k]$ *have* $p \ge 2$ *elements, and suppose the vectors* $(\boldsymbol{a}_i)_{i\in J}$ *in* $\mathbb{R}^n$ *satisfy*

$$K^{-3} \le \|\boldsymbol{a}_i\|_2 \le K^3 \quad (i \in J), \qquad \|\boldsymbol{a}_i - \boldsymbol{a}_{i'}\|_2 \ge K^{-3} \quad (i \ne i').$$

Fix $\varepsilon_{\mathrm{fam}} \ge 0$*. Suppose the symmetric matrices* $\widehat{D}_1, \dots, \widehat{D}_n \in \mathbb{R}^{p\times p}$ *form an* $\varepsilon_{\mathrm{fam}}$*-approximate observable-operator family for* $(\boldsymbol{a}_i)_{i\in J}$ *(Definition 6.4). Given these matrices and* $\delta \in (0, 1/3)$*,* RecoverDirections *(Algorithm 6.3) returns unit vectors* $\widehat{\boldsymbol{u}}_1, \dots, \widehat{\boldsymbol{u}}_p$ *such that, with probability at least* $1 - \delta$*, there is a bijection* $\pi : [p] \to J$ *satisfying*

$$\max_{i\in[p]} \left\| \widehat{\boldsymbol{u}}_i - \frac{\boldsymbol{a}_{\pi(i)}}{\|\boldsymbol{a}_{\pi(i)}\|_2} \right\|_2 \le 32K^6\sqrt{n}\, p^2 \varepsilon_{\mathrm{fam}}. \tag{44}$$

Moreover, the routine runs in time $\mathrm{poly}(n,k)\log(1/\delta)$.

Given an observable-operator family, the natural approach is to perform an eigendecomposition. Assuming a simple spectrum, this allows us to extract the eigenvectors and then the eigenvalues. We will use random contractions to ensure a sufficient eigenvalue gap.

Putting it Together (Section 6.4). Once we have analyzed the subroutines, it remains only to put them together and verify that their assumptions hold. The rest of this section presents the proofs of the results above.

6.2 Iterative Projections

Let $Q \subseteq [k]$ index the set of unrecovered regressors. The first step in the roadmap (Section 6.1) to analyzing our implicit tensor decomposition algorithm is an iterative projection procedure that computes estimates of the projections of restricted even and odd Hermite moments $A_Q^{(s)}, M_Q^{(s)}(b)$ (Theorem 6.3). This allows us to avoid explicitly materializing these high-order tensors. In this section, we present an iterative projection procedure inspired by Li and Liu [LL24] and Diakonikolas and Kane [DK25]. Starting from the trivial degree-1 projection map $C_1 = I_n$, we will show how to implicitly compute the degree-$(r+1)$ projection map C_{r+1} as well as the projected moment $C_{r+1}A_Q^{(r+1)}C_{r+1}^\top$, given the degree-$r$ projection map and projected moment. We will store an implicit representation of C_s and the projected odd moments can then be obtained given this representation.

We first describe the implicit representation in Section 6.2.1. Our algorithm is easiest to state assuming a moment estimation oracle EstimateMomentsByPeeling (Algorithm 8.1) that takes as input projected samples $(\boldsymbol{X}, C_r\widetilde{\boldsymbol{X}}^{\otimes r}, Y_{\max})$ and estimates the expectation of the projected moment. In Section 6.2.2, we describe the specifics of the oracle needed for this section and defer further details and its proof to Section 8. As our projection map is computed iteratively, we analyze the cumulative error in Section 6.2.3. In Section 6.2.4, we combine all the ingredients to prove the guarantees of our iterative projection procedure (Theorem 6.3).

6.2.1 Representation of Projection Maps

Fix a recovered list $\widehat{\mathcal{R}}$ matched bijectively to $R \subsetneq [k]$, and put $Q = [k] \setminus R$. Recall that $\boldsymbol{u}_i = \boldsymbol{w}_i / \|\boldsymbol{w}_i\|_2$, and $A_Q^{(r)} = \sum_{i\in Q} \lambda_i^{(2r)} \boldsymbol{u}_i^{\otimes r} (\boldsymbol{u}_i^{\otimes r})^\top$. We would like to construct a projection map $C_s \in \mathbb{R}^{p_s \times n^s}$ so that $\mathrm{range}(C_s^\top)$ approximately contains the lifted vectors $\boldsymbol{u}_i^{\otimes s}$ of components whose weights $\lambda_i^{(2s)}$ are sufficiently large relative to $\mathrm{tr}(A_Q^{(s)})$. We build C_s one tensor order at a time, starting with $C_1 = I_n$ and $p_1 = n$. For $2 \le r \le s$, put $p_r = \min\{|Q|, p_{r-1}n\}$. Since $\boldsymbol{u}_i^{\otimes r} = \boldsymbol{u}_i^{\otimes(r-1)} \otimes \boldsymbol{u}_i$,

$$\boldsymbol{u}_i^{\otimes(r-1)} \in \mathrm{range}(C_{r-1}^\top) \quad \Longrightarrow \quad \boldsymbol{u}_i^{\otimes r} \in \mathrm{range}(C_{r-1}^\top \otimes I_n).$$

For $r \ge 2$, assume that we have used EstimateMomentsByPeeling (Algorithm 8.1) to produce an intermediate estimate $\widehat{A}_r$ of the projected target moment matrix

$$(C_{r-1} \otimes I_n) A_Q^{(r)} (C_{r-1} \otimes I_n)^\top. \tag{45}$$

Note that the case of $C_1 = I_n$ is trivial. This target matrix has rank at most $|Q|$. Let the rows of V_r be the p_r leading orthonormal eigenvectors of $\widehat{A}_r$, and set

$$C_r := V_r (C_{r-1} \otimes I_n). \tag{46}$$

Our algorithm can proceed to the next iteration. Note that we only need to store $V_2, \ldots, V_s$.

Remark that we only used the even moments to determine the projection map. For the odd moment $M_Q^{(s)}(\boldsymbol{b})$, it calls EstimateMomentsByPeeling (Algorithm 8.1) once again to estimate all coordinate contractions

$$C_s M_Q^{(s)}(\boldsymbol{e}_\ell) C_s^\top \tag{47}$$

for $\ell \in [n]$. Since the odd moment is linear in the contraction direction, the matrices $\widehat{M}_1, \dots, \widehat{M}_n$ determine the family

$$\widehat{M}(\boldsymbol{b}) := \sum_{\ell=1}^{n} b_\ell \widehat{M}_\ell. \tag{48}$$

6.2.2 Mean Estimation Oracle

We now describe how to obtain estimates of (45) and (47) using EstimateMomentsByPeeling (Algorithm 8.1). To form $\widehat{A}_{r+1}$, we use the factors $V_2, \dots, V_r$. Once V_s is available, the same recursion gives the final odd features. For this purpose, we use the fact that the flattened Hermite moments can be written as expectations of outer products of the complexified covariates. For each observation, draw an independent $\boldsymbol{X}' \sim N(0, I_n)$ and put $\widetilde{\boldsymbol{X}} = \boldsymbol{X} + \mathrm{i}\boldsymbol{X}'$. By the representation of the projection map C_r (46), the projected powers satisfy the implicit tensor recursion

$$\begin{aligned} C_1 \widetilde{\boldsymbol{X}} &= \widetilde{\boldsymbol{X}}, \\ C_{r+1} \widetilde{\boldsymbol{X}}^{\otimes(r+1)} &= V_{r+1} \left[\left(C_r \widetilde{\boldsymbol{X}}^{\otimes r} \right) \otimes \widetilde{\boldsymbol{X}} \right]. \end{aligned} \tag{49}$$

Thus given $V_2, \dots, V_r$, we can evaluate $C_r \widetilde{\boldsymbol{X}}^{\otimes r}$ without forming $\widetilde{\boldsymbol{X}}^{\otimes r}$ or C_r. The recursion has at most s levels and uses tensor products of vectors and matrix-vector products in dimension at most n^2. Thus forming each projected sample takes time polynomial in s and n.

For each fresh observation, the even call at order $r+1$ uses $(C_r \widetilde{\boldsymbol{X}}^{\otimes r}) \otimes \widetilde{\boldsymbol{X}}$, and the final odd call uses $C_s \widetilde{\boldsymbol{X}}^{\otimes s}$. For $1 \le r < s$, the following even feature is used to estimate $\widehat{A}_{r+1}$, which determines V_{r+1}.

$$\Psi_{r+1} := \frac{\mathrm{Re}[\left((C_r \widetilde{\boldsymbol{X}}^{\otimes r}) \otimes \widetilde{\boldsymbol{X}} \right) \left((C_r \widetilde{\boldsymbol{X}}^{\otimes r}) \otimes \widetilde{\boldsymbol{X}} \right)^\top]}{\sqrt{(2r+2)!}}, \tag{50}$$

$$\mathbb{E}_{\boldsymbol{X}'}[\Psi_{r+1} \mid \boldsymbol{X}] = (C_r \otimes I_n)\, \mathrm{Mat}_{r+1}(\mathsf{H}_{2r+2}(\boldsymbol{X}))(C_r \otimes I_n)^\top, \tag{51}$$

$$\mathbb{E}_{\boldsymbol{X}, \xi_Q, \boldsymbol{X}'}[\Psi_{r+1} \mathbf{1}\{T \le Y_{\max,Q} \le T+1\}] \approx (C_r \otimes I_n) A_Q^{(r+1)} (C_r \otimes I_n)^\top.$$

At the final level, the odd features used to estimate $(\widehat{M}_\ell)_{\ell \in [n]}$ are

$$\Psi_{\mathrm{odd},\ell} := \frac{\mathrm{Re}[(C_s \widetilde{\boldsymbol{X}}^{\otimes s})(C_s \widetilde{\boldsymbol{X}}^{\otimes s})^\top \widetilde{X}_\ell]}{\sqrt{(2s+1)!}}, \qquad \ell \in [n], \tag{52}$$

$$\mathbb{E}_{\boldsymbol{X}'}[\Psi_{\mathrm{odd},\ell} \mid \boldsymbol{X}] = C_s\, \mathrm{Mat}_s(\mathsf{H}_{2s+1}(\boldsymbol{X}) \lrcorner \boldsymbol{e}_\ell) C_s^\top, \tag{53}$$

$$\mathbb{E}_{\boldsymbol{X}, \xi_Q, \boldsymbol{X}'}[\Psi_{\mathrm{odd},\ell} \mathbf{1}\{T \le Y_{\max,Q} \le T+1\}] \approx C_s M_Q^{(s)}(\boldsymbol{e}_\ell) C_s^\top.$$

Since $Y_{\max,Q}$ is not observed, EstimateMomentsByPeeling (Algorithm 8.1) uses the peeling weight $\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X}, Y_{\max})$ in place of the indicator. We give its estimation guarantee in Section 8.

6.2.3 Cumulative Projection Error

We now present the pseudocode of our iterative projection algorithm (Algorithm 6.1) and bound the cumulative projection error in terms of the accuracy of the intermediate moment estimates.

Lemma 6.7 (Recursive Projection Error). *Consider the Gaussian self-selection model (Definition 1.1) under the boundedness and uncoveredness assumptions (Assumption 1.2). Use the hyperparameters K, T, s from Hyperparameters 4.1. Suppose $Q = [k] \setminus R \neq \varnothing$ indexes the unrecovered regressors. Form $C_1, \dots, C_s$ by the intermediate eigenvector steps of* EstimateProjectedMoments *(Algorithm 6.1).*

For $2 \le r \le s$, let γ_r be the Frobenius error

$$\gamma_r := \left\| \widehat{A}_r - (C_{r-1} \otimes I_n) A_Q^{(r)} (C_{r-1} \otimes I_n)^\top \right\|_F .$$

For $1 \le t \le s$, put $\Pi_t = (C_t^\top C_t) \otimes I_{n^{s-t}}$.

(i) For every $1 \le t \le s$,

$$\left\| (I - \Pi_t)(A_Q^{(s)})^{1/2} \right\|_{\mathrm{op}}^2 \le 2 \sum_{r=2}^{t} \gamma_r \max_{i \in Q} \frac{\lambda_i^{(2s)}}{\lambda_i^{(2r)}}. \tag{54}$$

(ii) Fix $\varepsilon_{\mathrm{mean}} > 0$. If $\gamma_r \le \varepsilon_{\mathrm{mean}} \operatorname{tr}(A_Q^{(r)})$ for every $2 \le r \le s$, then, for every $1 \le t \le s$,

$$\left\| (I - \Pi_t)(A_Q^{(s)})^{1/2} \right\|_{\mathrm{op}}^2 \le 2\varepsilon_{\mathrm{mean}} K^{s-1} \operatorname{tr}(A_Q^{(s)}). \tag{55}$$

Algorithm 6.1: EstimateProjectedMoments

Input: The recovered list $\widehat{\mathcal{R}}$; sampling accuracy $\varepsilon_{\mathrm{est}} \in (0,1)$; failure probability $\delta \in (0, 1/3)$
Output: Symmetric $p_s \times p_s$ matrices $\widehat{A}, \widehat{M}_1, \dots, \widehat{M}_n$

$p_1 \leftarrow n$;
for $r = 1, \dots, s-1$ **do**
  // Algorithm 8.1 and Theorem 8.1(i)
  $\widehat{A}_{r+1} \leftarrow \texttt{EstimateMomentsByPeeling}(\widehat{\mathcal{R}}, \{V_2, \dots, V_r\}, \Psi_{r+1}, \varepsilon_{\mathrm{est}}, \delta/s)$;
  $p_{r+1} \leftarrow \min\{k - |\widehat{\mathcal{R}}|, p_r n\}$;
  Compute the eigenvalues $\nu_1 \ge \cdots \ge \nu_{p_r n}$ of $\widehat{A}_{r+1}$ and corresponding orthonormal eigenvectors $v_1, \dots, v_{p_r n}$ using a fixed deterministic convention;
  $V_{r+1} \leftarrow \begin{bmatrix} v_1^\top \\ \vdots \\ v_{p_{r+1}}^\top \end{bmatrix}$;
$\widehat{A} \leftarrow V_s \widehat{A}_s V_s^\top$;
// Algorithm 8.1 and Theorem 8.1(ii)
$(\widehat{M}_1, \dots, \widehat{M}_n) \leftarrow \texttt{EstimateMomentsByPeeling}(\widehat{\mathcal{R}}, \{V_2, \dots, V_r\}, (\Psi_{\mathrm{odd},\ell})_{\ell \in [n]}, \varepsilon_{\mathrm{est}}, \delta/s)$;
return $(\widehat{A}, \widehat{M}_1, \dots, \widehat{M}_n)$;

We are now ready to prove Lemma 6.7.

Proof. To prove Lemma 6.7(i), we bound the error at each projection, transfer the bounds to order s, and add them using the nested projection maps. For Lemma 6.7(ii), we compare the component weights at orders r and s.

Error at one projection. Fix $2 \le r \le s$. We compare the leading eigenspace of $\widehat{A}_r$ with the projected target moment matrix

$$L_r := (C_{r-1} \otimes I_n) A_Q^{(r)} (C_{r-1} \otimes I_n)^\top.$$

Let $P_r = V_r^\top V_r$, where the rows of V_r are the transposes of the p_r leading orthonormal eigenvectors of $\widehat{A}_r$. Since the component weights are positive, L_r is positive semidefinite with rank at most $|Q|$. By the dimension rule in Algorithm 6.1, $p_r = \min\{|Q|, p_{r-1}n\}$. If $P_r = I$, the following one-level bound is immediate. Otherwise $p_r = |Q|$ and $\lambda_{p_r+1}(L_r) = 0$. For $\boldsymbol{x} \in \operatorname{range}(I - P_r)$, Weyl's eigenvalue perturbation inequality (Proposition 4.6) gives

$$\begin{aligned}
\boldsymbol{x}^\top L_r \boldsymbol{x} &\le \boldsymbol{x}^\top \widehat{A}_r \boldsymbol{x} + \gamma_r \|\boldsymbol{x}\|_2^2 && (\left\|\widehat{A}_r - L_r\right\|_{\mathrm{op}} \le \gamma_r) \\
&\le (\lambda_{p_r+1}(\widehat{A}_r) + \gamma_r) \|\boldsymbol{x}\|_2^2 && (\boldsymbol{x} \in \operatorname{range}(I - P_r)) \\
&\le 2\gamma_r \|\boldsymbol{x}\|_2^2 && \text{(by Proposition 4.6)} \\
0 \preceq (I - P_r) L_r (I - P_r) &\preceq 2\gamma_r (I - P_r). && (\text{all } \boldsymbol{x} \in \operatorname{range}(I - P_r)) \quad (56)
\end{aligned}$$

The one-level bound (56) applies for $2 \le r \le s$. We still need to compare these bounds at order s.

Projection maps at order s. We now express the successive projection maps at order s. Their nested ranges will let us telescope the projection maps. Since C_{r-1} has orthonormal rows,

$$\begin{aligned}
(C_{r-1} \otimes I_n)(C_{r-1} \otimes I_n)^\top &= (C_{r-1} C_{r-1}^\top) \otimes I_n \\
&= I_{p_{r-1}n}. \qquad (57)
\end{aligned}$$

The recursion $C_r = V_r (C_{r-1} \otimes I_n)$ and $P_r = V_r^\top V_r$ therefore give

$$\begin{aligned}
\Pi_{r-1} &= (C_{r-1}^\top C_{r-1}) \otimes I_{n^{s-r+1}} \\
&= \left[(C_{r-1} \otimes I_n)^\top (C_{r-1} \otimes I_n)\right] \otimes I_{n^{s-r}} \\
\Pi_r &= (C_r^\top C_r) \otimes I_{n^{s-r}} \\
&= \left[(C_{r-1} \otimes I_n)^\top P_r (C_{r-1} \otimes I_n)\right] \otimes I_{n^{s-r}} \\
\Pi_{r-1}\Pi_r &= \left[(C_{r-1} \otimes I_n)^\top I_{p_{r-1}n} P_r (C_{r-1} \otimes I_n)\right] \otimes I_{n^{s-r}} && \text{(by (57))} \\
&= \Pi_r && (58) \\
D_r := \Pi_{r-1} - \Pi_r &= \left[(C_{r-1} \otimes I_n)^\top (I_{p_{r-1}n} - P_r)(C_{r-1} \otimes I_n)\right] \otimes I_{n^{s-r}}. && (59)
\end{aligned}$$

Since the Π_r are orthogonal projectors, (58) gives, for $2 \le r \le s$ and $1 \le t \le s$,

$$\begin{aligned}
D_r^2 &= (\Pi_{r-1} - \Pi_r)^2 \\
&= \Pi_{r-1} - 2\Pi_r + \Pi_r = D_r && \text{(by (58))} \\
I - \Pi_t &= \sum_{r=2}^{t} D_r. && (\Pi_1 = I; \text{ telescoping}) \quad (60)
\end{aligned}$$

Thus each D_r is an orthogonal projector, with $\|D_r\|_{\mathrm{op}} \le 1$. We next bound $D_r A_Q^{(s)} D_r$.

Error at order s. We first express (56) at order s. By (59) and the definition of L_r,

$$\begin{aligned}
D_r (A_Q^{(r)} \otimes I_{n^{s-r}}) D_r &= \left[(C_{r-1} \otimes I_n)^\top (I - P_r) L_r (I - P_r)(C_{r-1} \otimes I_n)\right] \otimes I_{n^{s-r}} \\
&\preceq 2\gamma_r D_r. && \text{(by (56))} \quad (61)
\end{aligned}$$

Since the $\boldsymbol{u}_i$ are unit vectors, $(\boldsymbol{u}_i\boldsymbol{u}_i^\top)^{\otimes(s-r)} \preceq I_{n^{s-r}}$. With zeroth tensor powers interpreted as 1, we obtain

$$\begin{aligned} A_Q^{(s)} &= \sum_{i\in Q} \lambda_i^{(2s)} \boldsymbol{u}_i^{\otimes r}(\boldsymbol{u}_i^{\otimes r})^\top \otimes (\boldsymbol{u}_i\boldsymbol{u}_i^\top)^{\otimes(s-r)} \\ &\preceq \sum_{i\in Q} \lambda_i^{(2s)} \boldsymbol{u}_i^{\otimes r}(\boldsymbol{u}_i^{\otimes r})^\top \otimes I_{n^{s-r}} && \text{(unit } \boldsymbol{u}_i\text{)} \\ &\preceq \left(\max_{i\in Q} \frac{\lambda_i^{(2s)}}{\lambda_i^{(2r)}}\right)(A_Q^{(r)} \otimes I_{n^{s-r}}). && \text{(positive component weights)} \end{aligned}$$

Conjugating by D_r, taking operator norms, and using $\|D_r\|_{\mathrm{op}} \le 1$ gives

$$\begin{aligned} \left\|D_r A_Q^{(s)} D_r\right\|_{\mathrm{op}} &\le \left(\max_{i\in Q} \frac{\lambda_i^{(2s)}}{\lambda_i^{(2r)}}\right)\left\|D_r(A_Q^{(r)} \otimes I_{n^{s-r}})D_r\right\|_{\mathrm{op}} && \text{(positive semidefinite order)} \\ &\le 2\gamma_r \max_{i\in Q} \frac{\lambda_i^{(2s)}}{\lambda_i^{(2r)}}. && \text{(by (61))} \quad (62) \end{aligned}$$

It remains to add these bounds over the levels.

Accumulating projection error. We now combine (62) using the telescoping identity (60). The Gram-matrix norm identity (Proposition A.4) gives

$$\begin{aligned} \left\|(I-\Pi_t)(A_Q^{(s)})^{1/2}\right\|_{\mathrm{op}}^2 &= \left\|(A_Q^{(s)})^{1/2}(I-\Pi_t)(A_Q^{(s)})^{1/2}\right\|_{\mathrm{op}} && \text{(by Proposition A.4)} \\ &= \left\|\sum_{r=2}^{t}(A_Q^{(s)})^{1/2} D_r (A_Q^{(s)})^{1/2}\right\|_{\mathrm{op}} && \text{(by (60))} \\ &\le \sum_{r=2}^{t}\left\|(A_Q^{(s)})^{1/2} D_r (A_Q^{(s)})^{1/2}\right\|_{\mathrm{op}} && \text{(triangle inequality)} \\ &= \sum_{r=2}^{t}\left\|D_r A_Q^{(s)} D_r\right\|_{\mathrm{op}} && \text{(by Proposition A.4)} \\ &\le 2\sum_{r=2}^{t}\gamma_r \max_{i\in Q}\frac{\lambda_i^{(2s)}}{\lambda_i^{(2r)}}. && \text{(by (62))} \end{aligned}$$

This proves Lemma 6.7(i). To prove Lemma 6.7(ii), we compare the trace at order r with the trace at order s. The component-weight comparison (Theorem 5.1(ii)) gives, for $i, j \in Q$, $\lambda_i^{(2s)}\lambda_j^{(2r)} \le K^{s-r}\lambda_i^{(2r)}\lambda_j^{(2s)}$. Summing over j and using the positive weights yields

$$\max_{i\in Q}\frac{\lambda_i^{(2s)}}{\lambda_i^{(2r)}} \le K^{s-r}\frac{\operatorname{tr}(A_Q^{(s)})}{\operatorname{tr}(A_Q^{(r)})}. \tag{63}$$

Substituting the assumed bounds on γ_r and (63) in Lemma 6.7(i) gives

$$\begin{aligned} \left\|(I-\Pi_t)(A_Q^{(s)})^{1/2}\right\|_{\mathrm{op}}^2 &\le 2\varepsilon_{\mathrm{mean}}\operatorname{tr}(A_Q^{(s)})\sum_{r=2}^{t}K^{s-r} && \text{(by (63))} \\ &\le 2\varepsilon_{\mathrm{mean}}K^{s-1}\operatorname{tr}(A_Q^{(s)}). && \text{(geometric sum, } K\ge 2\text{)} \end{aligned}$$

This concludes the proof. □

6.2.4 Proof of Theorem 6.3

We are now ready to prove the guarantees of our iterative projection algorithm (Algorithm 6.1). The proof calls on the estimation guarantees for EstimateMomentsByPeeling (Algorithm 8.1) that are presented in Section 8 as well as Lemma 6.7.

Proof of Theorem 6.3. For the estimation calls to EstimateMomentsByPeeling (Algorithm 8.1) for even moments at order $2 \le r \le s$, put $U = C_{r-1} \otimes I_n$. The even moment estimation guarantees in Theorem 8.1(i) and the restricted-moment decomposition in Theorem 5.1(i) give

$$\begin{aligned}
\gamma_r &= \left\| \widehat{A}_r - (C_{r-1} \otimes I_n) A_Q^{(r)} (C_{r-1} \otimes I_n)^\top \right\|_F \\
&\le \underbrace{\left\| \widehat{A}_r - U \operatorname{Mat}_r(\mathsf{T}_{2r}^Q) U^\top \right\|_F}_{\text{peeling estimate}} + \underbrace{\left\| U E_r U^\top \right\|_F}_{\text{restricted-moment remainder}} && \text{(by Theorem 5.1(i))} \\
&\le \underbrace{\left(1 + \sqrt{k/(2\pi)}\right) e^{-K^6 s} + k K^{4s} e^{-T^2/(2K)} \operatorname{tr}(A_Q^{(r)})}_{\text{by Theorem 8.1(i)}} \\
&\quad + \underbrace{K^{8s} e^{-T^2/(2K)} \operatorname{tr}(A_Q^{(r)})}_{\text{by Theorem 5.1(i) and } \|U\|_{\mathrm{op}} = 1} .
\end{aligned}$$

By Theorem 5.1(ii), $1 \le e^{T^2} \operatorname{tr}(A_Q^{(r)})$. Applying this bound to the additive term and simplifying the bound on γ_r using the choice of hyperparameters (Proposition D.1) gives

$$\gamma_r \le e^{-K^2 s} \operatorname{tr}(A_Q^{(r)}).$$

Applying the odd moment estimation guarantees from Theorem 8.1(ii) to the final call to EstimateMomentsByPeeling (Algorithm 8.1) for the odd moment and repeating the calculation at $r = s$ gives, simultaneously for every $\boldsymbol{b} \in \mathbb{R}^n$,

$$\left\| \widehat{M}(\boldsymbol{b}) - C_s M_Q^{(s)}(\boldsymbol{b}) C_s^\top \right\|_F \le e^{-K^2 s} \operatorname{tr}(A_Q^{(s)}) \, \|\boldsymbol{b}\|_2 .$$

Each call draws independent samples and satisfies its bound with probability at least $1 - \delta/s$. By a union bound over the $s - 1$ even calls and the one odd call, all bounds hold with probability at least $1 - \delta$. We proceed conditioning on the successful estimation.

For the final even estimate, take $U = C_{s-1} \otimes I_n$. Then

$$\begin{aligned}
\left\| \widehat{A} - C_s A_Q^{(s)} C_s^\top \right\|_F &= \left\| V_s \widehat{A}_s V_s^\top - V_s U A_Q^{(s)} U^\top V_s^\top \right\|_F && (\widehat{A} = V_s \widehat{A}_s V_s^\top, C_s = V_s U) \\
&= \left\| V_s \left(\widehat{A}_s - U A_Q^{(s)} U^\top \right) V_s^\top \right\|_F \\
&\le \left\| \widehat{A}_s - U A_Q^{(s)} U^\top \right\|_F && (\|V_s\|_{\mathrm{op}} = 1) \\
&\le e^{-K^2 s} \operatorname{tr}(A_Q^{(s)}). && \text{(even-call bound at } r = s)
\end{aligned}$$

Together with the odd-moment bound for every $\boldsymbol{b} \in \mathbb{R}^n$, this shows that the returned matrices $\widehat{A}, \widehat{M}_1, \ldots, \widehat{M}_n$ form an $e^{-K^2 s}$-estimate of the projected moment matrices for Q relative to the constructed map C_s (Definition 6.2). Moreover, by Lemma 6.7(ii), it follows that

$$\left\| (I - C_s^\top C_s)(A_Q^{(s)})^{1/2} \right\|_{\mathrm{op}}^2 \le 2K^{s-1} e^{-K^2 s} \operatorname{tr}(A_Q^{(s)}).$$

Sample complexity and running time. By Theorem 8.1, calling EstimateMomentsByPeeling (Algorithm 8.1) with p projected coordinates and m matrices uses

$$\widetilde{O}\left(p^2 m \, e^{2K^6 s} 100^s \log\left(\frac{s}{\delta}\right) \right)$$

projected samples and takes time at most the sample count times $\mathrm{poly}(n, k, p, m)$. The first even matrix has dimension n^2, and each later even matrix has dimension at most nk. The odd call estimates n matrices of dimension at most k. After simplifying by the choice of hyperparameters (Proposition D.1), this amounts to a total sample complexity of

$$\widetilde{O}\left((nk)^{4K^{10} + 12K^4} \log(1/\delta) \right). \tag{64}$$

Including the construction of the projected samples and the eigendecompositions gives total running time $(nk)^{\mathrm{poly}(K)} \log(1/\delta)$. □

6.3 Observable-Operator Learning

In Section 6.2, we constructed a projection map and estimated the even and odd moments in a lower-dimensional space. In this subsection, we use these estimates to recover a block of normalized regressors. On a high-level, we will construct an observable-operator family (Definition 6.4) whose eigenvalues encode projections $\langle \boldsymbol{b}, \boldsymbol{a}_i \rangle$ of the scaled regressors. Following Anandkumar, Hsu, and Kakade [AHK12], we diagonalize a random contraction and use its eigenvectors to read the coordinates of the scaled regressors.

6.3.1 Proof of Theorem 6.5

We first prove Theorem 6.5 by constructing an observable-operator family for a subset of the remaining scaled regressors. Let $Q \subseteq [k]$ index the unrecovered regressors and $\widehat{A}, \widehat{M}_1, \ldots, \widehat{M}_n$ denote the estimates to the projected moments returned by Algorithm 6.1. Roughly speaking, we have

$$\widehat{A} \approx \sum_{i \in Q} \lambda_i^{(2s)} (C_s \boldsymbol{u}_i^{\otimes s})(C_s \boldsymbol{u}_i^{\otimes s})^\top,$$
$$\sum_\ell b_\ell \widehat{M}_\ell \approx \sum_{i \in Q} \lambda_i^{(2s)} \langle \boldsymbol{b}, \boldsymbol{a}_i \rangle (C_s \boldsymbol{u}_i^{\otimes s})(C_s \boldsymbol{u}_i^{\otimes s})^\top.$$

We would like to rescale the odd moments by the even moment to cancel out the weights $\lambda_i^{(2s)}$

$$\widehat{A}^{-\frac{1}{2}} \left(\sum_\ell b_\ell \widehat{M}_\ell \right) \widehat{A}^{-\frac{1}{2}} \approx \sum_{i \in Q} \langle \boldsymbol{b}, \boldsymbol{a}_i \rangle (C_s \boldsymbol{u}_i^{\otimes s})(C_s \boldsymbol{u}_i^{\otimes s})^\top.$$

For this to be a well-conditioned operation, we would need the eigenvalues of $\widehat{A}$ to be bounded away from 0. Unfortunately, due to the irreducible bias that we swept under the rug and the fact that we only have approximate orthonormality, we cannot guarantee this is the case. Hence we first truncate $\widehat{A}$ to its leading eigenspace up to the largest gap between eigenvalues. Then we apply the rescaling procedure using the truncated even moment and construct an observable-operator family for the scaled regressors corresponding to the truncated eigenvectors. The pseudocode for this procedure can be found in Algorithm 6.2.

Algorithm 6.2: CONSTRUCTOBSERVABLEOPERATORS

Input: Symmetric matrices $\widehat{A}, \widehat{M}_1, \ldots, \widehat{M}_n \in \mathbb{R}^{p_s \times p_s}$
Output: A dimension $i_\star$ and matrices $\widehat{D}_1, \ldots, \widehat{D}_n$

Compute the eigenvalues $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_{p_s}$ and corresponding orthonormal eigenvectors $\widehat{v}_1, \ldots, \widehat{v}_{p_s}$ of $\widehat{A}$;
$\widehat{\lambda}_h \leftarrow \max\{\widehat{\lambda}_h, 0\}$ for $h \in [p_s]$, and $\widehat{\lambda}_{p_s+1} \leftarrow 0$;
$i_\star \leftarrow \arg\max_{1 \leq h \leq p_s} (\widehat{\lambda}_h - \widehat{\lambda}_{h+1})$;
$\widehat{V} \leftarrow \begin{bmatrix} \widehat{v}_1^\top \\ \vdots \\ \widehat{v}_{i_\star}^\top \end{bmatrix}$;
$\widehat{\Lambda} \leftarrow \operatorname{diag}(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{i_\star})$;
for $\ell = 1, \ldots, n$ **do**
  $\widehat{D}_\ell \leftarrow \widehat{\Lambda}^{-1/2} \widehat{V} \widehat{M}_\ell \widehat{V}^\top \widehat{\Lambda}^{-1/2}$;
return $(i_\star, \widehat{D}_1, \ldots, \widehat{D}_n)$;

Before proving Theorem 6.5, we state a useful spectral inequality whose proof is deferred to Appendix D.2.

Proposition 6.8 (Largest-Gap Eigenvalue Bounds). *Let $1 \leq p \leq q$, let $W \in \mathbb{R}^{p \times q}$, and let $\Lambda \in \mathbb{R}^{q \times q}$ be nonzero and positive semidefinite. Suppose the symmetric matrix $A \in \mathbb{R}^{p \times p}$ satisfies*

$$\left\| A - WW^\top \right\|_{\mathrm{op}} \leq \frac{\|\Lambda\|_{\mathrm{op}}}{8q}, \qquad \left\| W^\top W - \Lambda \right\|_{\mathrm{op}} \leq \frac{\|\Lambda\|_{\mathrm{op}}}{8q}.$$

Write the eigenvalues in nonincreasing order and put $\widehat{\lambda}_h = \max\{\lambda_h(A), 0\}, h \in [p]$, where we understand $\widehat{\lambda}_{p+1} = 0$ and $\lambda_{q+1}(\Lambda) = 0$. Then for the largest eigengap indexed by $i_\star \in \arg\max_{1 \leq h \leq p} (\widehat{\lambda}_h - \widehat{\lambda}_{h+1})$, we have

$$\widehat{\lambda}_{i_\star} \geq \frac{\|\Lambda\|_{\mathrm{op}}}{2q}, \qquad \widehat{\lambda}_{i_\star} - \lambda_{i_\star+1}(\Lambda) \geq \frac{\|\Lambda\|_{\mathrm{op}}}{2q}.$$

We are now ready to analyze Algorithm 6.2. The proof proceeds by decomposing the error into three interpretable terms. To avoid cumbersome calculations that distract from our understanding, we defer the analysis of the ones that are a consequence of the near orthogonality of the rank-one tensors to the appendix.

Proof of Theorem 6.5. We first describe the error decomposition assuming that $\widehat{\Lambda}$ is positive definite.

We verify positivity in the perturbation bounds below. Since $|Q| \le k$, the error assumption also holds with $|Q|$ in place of k.

Order the components by decreasing $\lambda_i^{(2s)}$. Define the factor matrix

$$H := \left[\sqrt{\lambda_i^{(2s)}}\, \boldsymbol{u}_i^{\otimes s}\right]_{i\in Q}, \qquad \Lambda := \operatorname{diag}_{i\in Q}(\lambda_i^{(2s)}),$$

so that

$$A_Q^{(s)} = HH^\top, \qquad M_Q^{(s)}(\boldsymbol{b}) = H \operatorname{diag}_{i\in Q}(\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle) H^\top.$$

Now, let $\widehat{V}^\top := [\widehat{\boldsymbol{v}}_1, \dots, \widehat{\boldsymbol{v}}_{i_\star}]$ be the matrix of top $i_\star$ eigenvectors of $\widehat{A}$, as chosen in Algorithm 6.2. Define the ideal projected, truncated, and rescaled factor matrix

$$F = [\boldsymbol{f}_i]_{i\in Q} = \widehat{\Lambda}^{-1/2}\widehat{V} C_s H, \qquad \boldsymbol{f}_i := \sqrt{\lambda_i^{(2s)}}\widehat{\Lambda}^{-1/2}\widehat{V}(C_s \boldsymbol{u}_i^{\otimes s}). \tag{65}$$

Note that by Theorem 5.1(ii), $\lambda_i^{(2s)} > 0$ so that $\Lambda \neq 0$ and its inverse square root is well-defined.

Let $J \subseteq Q$ index the first $i_\star$ components in the fixed order, and write F_J and $F_{Q\setminus J}$ for the corresponding column submatrices. Expanding into outer products gives

$$\begin{aligned} F \operatorname{diag}_{i\in Q}(\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle) F^\top &= \sum_{i\in Q} \langle \boldsymbol{b}, \boldsymbol{a}_i\rangle\, \boldsymbol{f}_i \boldsymbol{f}_i^\top \\ &= \sum_{i\in J} \langle \boldsymbol{b}, \boldsymbol{a}_i\rangle\, \boldsymbol{f}_i \boldsymbol{f}_i^\top + \sum_{i\in Q\setminus J} \langle \boldsymbol{b}, \boldsymbol{a}_i\rangle\, \boldsymbol{f}_i \boldsymbol{f}_i^\top \\ &= F_J \operatorname{diag}_{i\in J}(\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle) F_J^\top + F_{Q\setminus J} \operatorname{diag}_{i\in Q\setminus J}(\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle) F_{Q\setminus J}^\top. \end{aligned}$$

We will construct an orthogonal matrix $Z = [\boldsymbol{z}_i]_{i\in J}$. For $\widehat{D}(\boldsymbol{b}) = \sum_{\ell=1}^n b_\ell \widehat{D}_\ell$, the corresponding error decomposes as

$$\begin{aligned} &\widehat{D}(\boldsymbol{b}) - \sum_{i\in J} \langle \boldsymbol{b}, \boldsymbol{a}_i\rangle\, \boldsymbol{z}_i \boldsymbol{z}_i^\top \\ &\quad = \underbrace{\widehat{D}(\boldsymbol{b}) - F \operatorname{diag}_{i\in Q}(\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle) F^\top}_{\text{odd estimation error}} \\ &\qquad + \underbrace{F_{Q\setminus J} \operatorname{diag}_{i\in Q\setminus J}(\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle) F_{Q\setminus J}^\top}_{\text{spectral truncation error}} \\ &\qquad + \underbrace{F_J \operatorname{diag}_{i\in J}(\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle) F_J^\top - \sum_{i\in J} \langle \boldsymbol{b}, \boldsymbol{a}_i\rangle\, \boldsymbol{z}_i \boldsymbol{z}_i^\top}_{\text{approximate orthonormality}}. \end{aligned} \tag{66}$$

By the triangle inequality, it suffices to bound the Frobenius norm of each term.

The first term comes from estimating the projected odd moment. Rescaling can blow up this estimation error by at most a factor of $1/\widehat{\lambda}_{i_\star}$ in Frobenius norm. We therefore need a lower bound on the smallest eigenvalue used for rescaling.

For the second term, suppose the projected lifted vectors are orthonormal and the even moment is known exactly. Then the rows of $\widehat{V}$ span the projected directions indexed by J, so $\widehat{V} C_s \boldsymbol{u}_i^{\otimes s} = 0$ for

every $i \in Q \setminus J$. For the estimated even moment, we apply Sylvester's equation (Proposition D.3) and the eigenvalue gap to bound the contribution from these components.

In the same exact case, rescaling cancels the factor $\sqrt{\lambda_i^{(2s)}}$ in each column of F_J. The resulting vectors already form an orthonormal basis, so the third term vanishes. When the even moment is accurate on the block and little weight enters from outside J, only a small correction is needed to make the rescaled vectors orthonormal.

Perturbation bounds. We first compare the estimated even moment with the projected moment matrix, and the projected component matrix with the component weights. These bounds will verify the hypotheses of Proposition 6.8. By the even-moment estimation assumption (Definition 6.2(i)),

$$\begin{aligned}\left\|\widehat{A} - C_s A_Q^{(s)} C_s^\top\right\|_F &\le \varepsilon_{\text{mom}} \operatorname{tr}(A_Q^{(s)}) && \text{(by Definition 6.2(i))}\\ &\le |Q|\varepsilon_{\text{mom}} \|\Lambda\|_{\text{op}}. && (\operatorname{tr}(A_Q^{(s)}) = \operatorname{tr}(\Lambda) \le |Q|\,\|\Lambda\|_{\text{op}}) \quad (67)\end{aligned}$$

The second comparison separates approximate orthogonality from the change caused by projection. The approximate-orthogonality hypothesis (Assumption 6.5(b)) gives

$$\begin{aligned}\left\|H^\top H - \Lambda\right\|_{\text{op}} &= \left\|\Lambda^{1/2}\left[\left(\langle \boldsymbol{u}_i^{\otimes s}, \boldsymbol{u}_j^{\otimes s}\rangle\right)_{i,j\in Q} - I\right]\Lambda^{1/2}\right\|_{\text{op}}\\ &\le \varepsilon_\perp \|\Lambda\|_{\text{op}}. && \text{(by Assumption 6.5(b))} \quad (68)\end{aligned}$$

Using the orthonormality of the rows of C_s and the norm identity (Proposition A.4), we obtain

$$\begin{aligned}\left\|H^\top C_s^\top C_s H - \Lambda\right\|_{\text{op}} \le\ & \underbrace{\left\|H^\top H - \Lambda\right\|_{\text{op}}}_{\text{approximate orthogonality}}\\ &+ \underbrace{\left\|H^\top (I - C_s^\top C_s) H\right\|_{\text{op}}}_{\text{projection error}} && \text{(triangle inequality)}\\ &\le \varepsilon_\perp \|\Lambda\|_{\text{op}} + \left\|(I - C_s^\top C_s)H\right\|_{\text{op}}^2 && \text{(by (68) and Proposition A.4)}\\ &= \varepsilon_\perp \|\Lambda\|_{\text{op}}\\ &\quad + \left\|(I - C_s^\top C_s)(A_Q^{(s)})^{1/2}\right\|_{\text{op}}^2 && \text{(by Proposition A.4)}\\ &\le \varepsilon_\perp \|\Lambda\|_{\text{op}} + \varepsilon_{\text{proj}} \operatorname{tr}(A_Q^{(s)}) && \text{(by Assumption 6.5(d))}\\ &\le \left(|Q|\varepsilon_{\text{proj}} + \varepsilon_\perp\right)\|\Lambda\|_{\text{op}}. && (69)\end{aligned}$$

The bounds (67) and (69) verify the hypotheses of Proposition 6.8, since

$$\begin{aligned}|Q|\varepsilon_{\text{mom}} + |Q|\varepsilon_{\text{proj}} + \varepsilon_\perp &\le |Q|\left(\varepsilon_{\text{mom}} + \varepsilon_{\text{proj}} + \varepsilon_\perp\right)\\ &\le \frac{1}{16|Q|}. && \text{(by Assumption 6.5(e), } |Q| \le k)\end{aligned}$$

Each coefficient is therefore at most $1/(8|Q|)$, and $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_F$ converts the even estimation bound to operator norm. Applying Proposition 6.8 with $W = C_s H$ gives

$$\widehat{\lambda}_{i_\star} \ge \frac{\|\Lambda\|_{\mathrm{op}}}{2|Q|}. \tag{70}$$

This bound shows that $\widehat{\Lambda}^{-1/2}$ is defined and controls its norm.

Odd estimation error. We first bound the odd estimation error in (66). Using the definitions of $\widehat{D}(\boldsymbol{b})$ and F and the odd-moment estimation assumption (Definition 6.2(ii)), we obtain

$$\begin{aligned}
&\left\|\widehat{D}(\boldsymbol{b}) - F\,\mathrm{diag}_{i\in Q}(\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle)F^\top\right\|_F \\
&= \left\|\widehat{\Lambda}^{-1/2}\widehat{V}\left(\sum_{\ell=1}^{n} b_\ell \widehat{M}_\ell - C_s M_Q^{(s)}(\boldsymbol{b})C_s^\top\right)\widehat{V}^\top\widehat{\Lambda}^{-1/2}\right\|_F && \text{(by Algorithm 6.2 and (65))} \\
&\le \frac{\left\|\sum_{\ell=1}^{n} b_\ell \widehat{M}_\ell - C_s M_Q^{(s)}(\boldsymbol{b})C_s^\top\right\|_F}{\widehat{\lambda}_{i_\star}} \\
&\le \frac{\varepsilon_{\mathrm{mom}}\,\mathrm{tr}(A_Q^{(s)})}{\widehat{\lambda}_{i_\star}}\,\|\boldsymbol{b}\|_2 && \text{(by assumption (Definition 6.2(ii)))} \\
&\le 2|Q|^2\varepsilon_{\mathrm{mom}}\,\|\boldsymbol{b}\|_2\,. && \text{(by (70), } \mathrm{tr}(\Lambda) \le |Q|\,\|\Lambda\|_{\mathrm{op}}) \quad (71)
\end{aligned}$$

This bounds the first term in the error decomposition (66). For the last two terms, take Z to be an orthogonal polar factor of F_J. We defer the bounds for the spectral truncation error (Lemma D.4(ii)) and the approximate orthonormality term Lemma D.4(iii) to the appendix.

Putting it together. Applying the triangle inequality to (66) and substituting the three bounds gives

$$\begin{aligned}
&\left\|\widehat{D}(\boldsymbol{b}) - \sum_{i\in J}\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle\, \boldsymbol{z}_i\boldsymbol{z}_i^\top\right\|_F \\
&\le \Bigg(\underbrace{2|Q|^2\varepsilon_{\mathrm{mom}}}_{\text{odd estimation error}} && \text{(by (71))} \\
&\quad + \underbrace{K^3\left\|F_{Q\setminus J}\right\|_F^2}_{\text{spectral truncation error}} && \text{(by Lemma D.4(ii))} \\
&\quad + \underbrace{3K^3\left(2|Q|^2\varepsilon_{\mathrm{mom}} + \left\|F_{Q\setminus J}\right\|_F^2\right)}_{\text{approximate orthonormality}}\Bigg)\,\|\boldsymbol{b}\|_2 && \text{(by Lemma D.4(iii))} \\
&\le \left(8K^3|Q|^2\varepsilon_{\mathrm{mom}} + 4K^3\left\|F_{Q\setminus J}\right\|_F^2\right)\|\boldsymbol{b}\|_2 && (K \ge 1) \quad (72) \\
&\le \left(8K^3|Q|^2\varepsilon_{\mathrm{mom}} + 4K^3|Q|^3(\varepsilon_{\mathrm{mom}} + \varepsilon_{\mathrm{proj}} + \varepsilon_\perp)\right)\|\boldsymbol{b}\|_2 && \text{(by Lemma D.4(i))} \\
&\le 16K^3|Q|^3\left(\varepsilon_{\mathrm{mom}} + \varepsilon_{\mathrm{proj}} + \varepsilon_\perp\right)\|\boldsymbol{b}\|_2\,. && (|Q| \ge 1) \quad (73)
\end{aligned}$$

The same basis works for every $\boldsymbol{b} \in \mathbb{R}^n$. Since $|Q| \le k$, (73) gives the claimed family approximation in the sense of Definition 6.4. All matrices used by the routine have dimension at most $p_s \le |Q| \le k$, so it runs in $\mathrm{poly}(n, k)$ time. □

6.3.2 Proof of Theorem 6.6

Theorem 6.5 gives an approximate observable-operator family for a block $(\boldsymbol{a}_i)_{i\in J}$ that approximates

$$D(\boldsymbol{b}) = \sum_{i\in J} \langle \boldsymbol{b}, \boldsymbol{a}_i \rangle \, \boldsymbol{z}_i \boldsymbol{z}_i^\top.$$

Here, we prove Theorem 6.6, by demonstrating how to recover the normalized regressors given this family.

Observe that the eigenvalues of $D(\boldsymbol{e}_j)$ are the j-th coordinates of the vectors $\boldsymbol{a}_i$. A first attempt might be to compute the eigenvalues for each $D(\boldsymbol{e}_j)$ and assemble the vectors. However, an eigendecomposition outputs each coordinate as an unordered list, and does not specify the mapping to the particular regressor.

In order to match the coordinates to regressors, we must also recover the common eigenvectors $\boldsymbol{z}_i$'s. This can be accomplished via an eigendecomposition of $D(\boldsymbol{b})$, assuming its spectrum is simple. We can ensure a quantitative separation between the eigenvalues with constant probability by picking a random unit vector $\boldsymbol{b}$. We then recover the coordinates through

$$\boldsymbol{z}_i^\top D(\boldsymbol{e}_j) \boldsymbol{z}_i = (\boldsymbol{a}_i)_j \qquad (i \in J,\ j \in [n]).$$

The pseudocode for this subroutine can be found in Algorithm 6.3.

Algorithm 6.3: RECOVERDIRECTIONS

Input: Symmetric matrices $(\widehat{D}_j)_{j\in[n]}$ of common dimension $p \geq 2$, and confidence parameter δ
Output: Unit vectors $\widehat{\boldsymbol{u}}_1, \ldots, \widehat{\boldsymbol{u}}_p$

$L \leftarrow \log_2(1/\delta)$;
Draw independent $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_L \sim \mathrm{Unif}(\mathbb{S}^{n-1})$;
for $\iota = 1, \ldots, L$ **do**
  $\widehat{D}(\boldsymbol{b}_\iota) \leftarrow \sum_{j=1}^n (\boldsymbol{b}_\iota)_j \widehat{D}_j$;
  Compute the eigenvalues $\widehat{\nu}_1^{(\iota)} \geq \cdots \geq \widehat{\nu}_p^{(\iota)}$ and corresponding orthonormal eigenvectors $\widehat{\boldsymbol{z}}_1^{(\iota)}, \ldots, \widehat{\boldsymbol{z}}_p^{(\iota)}$ of $\widehat{D}(\boldsymbol{b}_\iota)$;
$\iota_\star \leftarrow \arg\max_{1\leq\iota\leq L} \min_{1\leq i<p} (\widehat{\nu}_i^{(\iota)} - \widehat{\nu}_{i+1}^{(\iota)})$;
For $i \in [p]$ and $j \in [n]$, set $(\widehat{\boldsymbol{a}}_i)_j \leftarrow (\widehat{\boldsymbol{z}}_i^{(\iota_\star)})^\top \widehat{D}_j \widehat{\boldsymbol{z}}_i^{(\iota_\star)}$;
$\widehat{\boldsymbol{u}}_i \leftarrow \widehat{\boldsymbol{a}}_i / \|\widehat{\boldsymbol{a}}_i\|_2$ for $i \in [p]$;
return $(\widehat{\boldsymbol{u}}_i)_{i\in[p]}$;

We now prove that $O(\log(1/\delta))$ random unit directions suffice to recover every normalized vector $\boldsymbol{a}_i / \|\boldsymbol{a}_i\|_2$ with bounded error.

Proof of Theorem 6.6. We first show that our algorithm finds an approximate common eigenbasis for the family. We then show that evaluating the coordinate quadratic forms in this basis and normalizing the resulting vectors gives the claimed error guarantees.

Distances between unit vectors are at most 2, so it suffices to prove the claim when

$$\varepsilon_{\mathrm{fam}} < \frac{1}{16K^6\sqrt{n}\,p^2}. \tag{74}$$

Recovering the common eigenvectors. Our strategy is to apply a Davis–Kahan type eigenvector bound (Proposition D.6) to control the eigenvector error by the operator-norm error divided by an eigenvalue gap. For unit contractions, the operator-norm error is bounded above by $\varepsilon_{\mathrm{fam}}$ by assumption, so we first establish the required eigenvalue gap.

Let $(z_i)_{i\in J}$ be the orthonormal basis supplied by assumption (Definition 6.4), and put

$$D(b) = \sum_{i\in J} \langle b, a_i\rangle\, z_i z_i^\top.$$

Its eigenvalues are $(\langle b, a_i\rangle)_{i\in J}$. For a uniformly random unit vector $b \sim \mathbb{S}^{n-1}$, spherical anti-concentration (Proposition D.5) and a union bound give

$$\begin{aligned}
&\mathbb{P}_b\left(\min_{i\neq i'\in J} |\langle b, a_i - a_{i'}\rangle| \le \frac{1}{K^3\sqrt{n}\,p^2}\right) \\
&\quad\le \sum_{\substack{i,i'\in J\\ i<i'}} \frac{1}{K^3 p^2\,\|a_i - a_{i'}\|_2} && \text{(union bound and Proposition D.5)} \\
&\quad\le \frac{\binom{p}{2}}{p^2} < \frac{1}{2}. && (\|a_i - a_{i'}\|_2 \ge K^{-3})
\end{aligned}$$

The probability that none of the $L = \log_2(1/\delta)$ independent trials has this separation is at most $2^{-L} \le \delta$. For the rest of the proof, suppose that at least one trial has this separation.

We now show that the eigendecomposition chosen by our algorithm approximates the common eigenbasis. By assumption, the family approximation bound and $\|b_\iota\|_2 = 1$ give

$$\left\|\widehat{D}(b_\iota) - D(b_\iota)\right\|_{\mathrm{op}} \le \left\|\widehat{D}(b_\iota) - D(b_\iota)\right\|_F \le \varepsilon_{\mathrm{fam}} \qquad (\iota \in [L]). \tag{75}$$

By Weyl's inequality (Proposition 4.6), each estimated consecutive eigengap differs from the corresponding true eigengap by at most $2\varepsilon_{\mathrm{fam}}$. Since our algorithm maximizes the minimum estimated gap,

$$\begin{aligned}
\min_{1\le i<p}\left(\widehat{v}_i^{(\iota_\star)} - \widehat{v}_{i+1}^{(\iota_\star)}\right) &\ge \frac{1}{K^3\sqrt{n}\,p^2} - 2\varepsilon_{\mathrm{fam}} && \text{(separated trial and choice of } \iota_\star) \\
&\ge \frac{1}{K^3\sqrt{n}\,p^2} - \frac{1}{8K^6\sqrt{n}\,p^2} && \text{(by (74))} \\
&\ge \frac{3}{4K^3\sqrt{n}\,p^2}. && (K \ge 1) \quad (76)
\end{aligned}$$

As this is at least $3/(4K^3\sqrt{n}\,p^2) > 2\varepsilon_{\mathrm{fam}}$ by (74), Weyl's inequality shows that $\widehat{D}(b_{\iota_\star})$ also has simple spectrum. Matching the true eigenvectors in eigenvalue order gives a bijection $\pi : [p] \to J$. Write $\widehat{z}_i = \widehat{z}_i^{(\iota_\star)}$ for the eigenvectors used by the algorithm. With $\widehat{D}(b_{\iota_\star})$ as the reference matrix, the

Davis–Kahan eigenvector bound gives, after choosing the signs of the true eigenvectors,

$$\begin{aligned}
\left\|\widehat{\boldsymbol{z}}_i - \boldsymbol{z}_{\pi(i)}\right\|_2 &\le \frac{2\sqrt{2}\,\left\|\widehat{D}(\boldsymbol{b}_{\iota_\star}) - D(\boldsymbol{b}_{\iota_\star})\right\|_{\mathrm{op}}}{\min_{1\le i'<p}\left(\widehat{v}_{i'}^{(\iota_\star)} - \widehat{v}_{i'+1}^{(\iota_\star)}\right)} && \text{(by Proposition D.6)} \\
&\le 2\sqrt{2}\,\varepsilon_{\mathrm{fam}}\frac{4K^3\sqrt{n}\,p^2}{3} && \text{(by (75) and (76))} \\
&\le 4K^3\sqrt{n}\,p^2\varepsilon_{\mathrm{fam}} && (2\sqrt{2}\le 3) \quad (77) \\
&< \frac{1}{4K^3}. && \text{(by (74))} \quad (78)
\end{aligned}$$

Recovering the normalized regressors. We now use the approximate common eigenvectors to estimate the scaled regressors. Since $\boldsymbol{z}_{\pi(i)}$ is an eigenvector of every $D(\boldsymbol{b})$ and both vectors are unit, the coordinate construction gives

$$\begin{aligned}
\left\langle \boldsymbol{b}, \widehat{\boldsymbol{a}}_i - \boldsymbol{a}_{\pi(i)}\right\rangle &= \widehat{\boldsymbol{z}}_i^\top\left(\widehat{D}(\boldsymbol{b}) - D(\boldsymbol{b})\right)\widehat{\boldsymbol{z}}_i \\
&\quad + (\widehat{\boldsymbol{z}}_i - \boldsymbol{z}_{\pi(i)})^\top\left(D(\boldsymbol{b}) - \left\langle \boldsymbol{b}, \boldsymbol{a}_{\pi(i)}\right\rangle I\right)(\widehat{\boldsymbol{z}}_i - \boldsymbol{z}_{\pi(i)}). \qquad (79)
\end{aligned}$$

For every unit vector $\boldsymbol{b}$,

$$\begin{aligned}
\left\|D(\boldsymbol{b}) - \left\langle \boldsymbol{b}, \boldsymbol{a}_{\pi(i)}\right\rangle I\right\|_{\mathrm{op}} &= \max_{j\in J}\left|\left\langle \boldsymbol{b}, \boldsymbol{a}_j - \boldsymbol{a}_{\pi(i)}\right\rangle\right| && \text{(common orthonormal eigenbasis)} \\
&\le \max_{j\in J}\left(\|\boldsymbol{a}_j\|_2 + \left\|\boldsymbol{a}_{\pi(i)}\right\|_2\right) && \text{(Cauchy–Schwarz and triangle inequality)} \\
&\le 2K^3. && (\|\boldsymbol{a}_j\|_2 \le K^3) \quad (80)
\end{aligned}$$

Taking the supremum over unit $\boldsymbol{b}$ in (79) and using the family approximation assumption (Definition 6.4) gives

$$\begin{aligned}
\left\|\widehat{\boldsymbol{a}}_i - \boldsymbol{a}_{\pi(i)}\right\|_2 &= \sup_{\|\boldsymbol{b}\|_2=1}\left|\left\langle \boldsymbol{b}, \widehat{\boldsymbol{a}}_i - \boldsymbol{a}_{\pi(i)}\right\rangle\right| \\
&\le \varepsilon_{\mathrm{fam}} + 2K^3\left\|\widehat{\boldsymbol{z}}_i - \boldsymbol{z}_{\pi(i)}\right\|_2^2 && \text{(by (80))} \\
&\le \varepsilon_{\mathrm{fam}} + \frac{1}{2}\left\|\widehat{\boldsymbol{z}}_i - \boldsymbol{z}_{\pi(i)}\right\|_2 && \text{(by (78))} \\
&\le \left(1 + 2K^3\sqrt{n}\,p^2\right)\varepsilon_{\mathrm{fam}} && \text{(by (77))} \\
&\le 3K^3\sqrt{n}\,p^2\varepsilon_{\mathrm{fam}}. && (K, n, p \ge 1) \quad (81)
\end{aligned}$$

By (81) and (74),

$$\left\|\widehat{\boldsymbol{a}}_i - \boldsymbol{a}_{\pi(i)}\right\|_2 < \frac{3}{16K^3} < K^{-3} \le \left\|\boldsymbol{a}_{\pi(i)}\right\|_2.$$

Thus every $\widehat{a}_i$ is nonzero. Since $\widehat{a}_i = \|\widehat{a}_i\|_2 \widehat{u}_i$ and $\|\widehat{u}_i\|_2 = 1$, normalization gives

$$\begin{aligned}
&\left\| \widehat{u}_i - \frac{a_{\pi(i)}}{\left\|a_{\pi(i)}\right\|_2} \right\|_2 \\
&= \left\| \frac{\widehat{a}_i - a_{\pi(i)}}{\left\|a_{\pi(i)}\right\|_2} + \left(1 - \frac{\|\widehat{a}_i\|_2}{\left\|a_{\pi(i)}\right\|_2}\right) \widehat{u}_i \right\|_2 \\
&\le \frac{\left\|\widehat{a}_i - a_{\pi(i)}\right\|_2 + \left| \left\|a_{\pi(i)}\right\|_2 - \|\widehat{a}_i\|_2 \right|}{\left\|a_{\pi(i)}\right\|_2} && \text{(triangle inequality)} \\
&\le \frac{2\left\|\widehat{a}_i - a_{\pi(i)}\right\|_2}{\left\|a_{\pi(i)}\right\|_2} && \text{(reverse triangle inequality)} \\
&\le 6K^6\sqrt{n}\, p^2 \varepsilon_{\mathrm{fam}}. && \text{(by (81); } \left\|a_{\pi(i)}\right\|_2 \ge K^{-3}) \quad (82)
\end{aligned}$$

This gives the claimed error bound for normalized vectors. Each of the $L = O(\log(1/\delta))$ trials forms an $p \times p$ matrix and computes its eigendecomposition. Since $p \le k$, these computations and the coordinate quadratic forms take $\mathrm{poly}(n,k)\log(1/\delta)$ time. □

6.4 Proof of Theorem 6.1

Having analyzed each of the subroutines in our implicit tensor decomposition algorithm (Algorithm 6.4), we are now ready to put them together and prove its guarantees. The pseudocode can be found in Algorithm 6.4.

Algorithm 6.4: IMPLICITTENSORDECOMPOSITION

Input: The recovered list $\widehat{\mathcal{R}}$ and failure probability δ

Output: A non-empty list of unit vectors $(\widehat{u}_1, \ldots, \widehat{u}_{i_\star})$ giving rough estimates of normalized regressors

```
// Algorithm 6.1 and Theorem 6.3
(Â, M̂_1, ..., M̂_n) ← EstimateProjectedMoments(R̂, e^{-K^6 s}, δ/2);
// Algorithm 6.2 and Theorem 6.5
(i_⋆, D̂_1, ..., D̂_n) ← ConstructObservableOperators(Â, (M̂_ℓ)_{ℓ∈[n]});
if i_⋆ = 1 then
    â_1 ← ((D̂_ℓ)_11)_{ℓ∈[n]};
    return (â_1 / ‖â_1‖_2);
// Algorithm 6.3 and Theorem 6.6
(û_1, ..., û_{i_⋆}) ← RecoverDirections((D̂_ℓ)_{ℓ∈[n]}, δ/2);
return (û_1, ..., û_{i_⋆});
```

We use Theorem 6.3 to estimate the projected moments and Theorem 6.5 to construct an observable-operator family for a non-empty block. For a singleton block we read the coordinates directly. For a larger block we apply Theorem 6.6 to recover the normalized regressors.

Proof of Theorem 6.1. We verify that the conditions to the respective results are satisfied.

Projected moments and projection error. In order to apply Theorem 6.3, we must check the accuracy of the recovered list and the inputs to EstimateProjectedMoments.

By hypothesis, the recovered regressors have estimation error at most $e^{-K^6 s}$. Hence Theorem 6.3 applies. Our algorithm calls EstimateProjectedMoments with sampling accuracy $e^{-K^6 s}$ and failure probability $\delta/2$, using fresh observations. As a result, with probability at least $1 - \delta/2$, the following holds, where $\varepsilon_{\mathrm{mean}} = e^{-K^2 s}$:

$$\begin{aligned}
\left\| \widehat{A} - C_s A_Q^{(s)} C_s^\top \right\|_F &\le \varepsilon_{\mathrm{mean}} \operatorname{tr}(A_Q^{(s)}), \\
\left\| \sum_{\ell=1}^n b_\ell \widehat{M}_\ell - C_s M_Q^{(s)}(\boldsymbol{b}) C_s^\top \right\|_F &\le \varepsilon_{\mathrm{mean}} \operatorname{tr}(A_Q^{(s)}) \, \|\boldsymbol{b}\|_2 , \\
\left\| (I - C_s^\top C_s)(A_Q^{(s)})^{1/2} \right\|_{\mathrm{op}}^2 &\le 2\varepsilon_{\mathrm{mean}} K^{s-1} \operatorname{tr}(A_Q^{(s)}).
\end{aligned}$$

The odd-moment bound holds simultaneously for every $\boldsymbol{b} \in \mathbb{R}^n$. We proceed assuming fixed estimates that satisfy these bounds.

Constructing the observable-operator family. In order to apply Theorem 6.5, we must check the moment and projection errors, approximate orthogonality of the lifted vectors, the upper bound on the scaled-regressor norms, and the total error bound.

By construction, $C_s \in \mathbb{R}^{p_s \times n^s}$ has orthonormal rows and $1 \le p_s \le |Q|$. The preceding estimates give

$$\begin{aligned}
\varepsilon_{\mathrm{mom}} &= \varepsilon_{\mathrm{mean}}, \\
\varepsilon_{\mathrm{proj}} &= 2K^{s-1}\varepsilon_{\mathrm{mean}}.
\end{aligned}$$

The approximate-orthogonality bound (Theorem 5.1(iii)) gives the required estimate with $\varepsilon_\perp = ke^{-s/K}$:

$$\begin{aligned}
\left\| \left(\langle \boldsymbol{u}_i^{\otimes s}, \boldsymbol{u}_j^{\otimes s} \rangle \right)_{i,j \in Q} - I \right\|_{\mathrm{op}} &\le \max_{i \in Q} \sum_{j \in Q \setminus \{i\}} \left| \langle \boldsymbol{u}_i^{\otimes s}, \boldsymbol{u}_j^{\otimes s} \rangle \right| && \text{(maximum absolute row sum)} \\
&\le (|Q| - 1) e^{-s/K} \le k e^{-s/K} = \varepsilon_\perp . && \text{(by Theorem 5.1(iii))}
\end{aligned}$$

By mundane calculations deferred to the appendix (Proposition C.1),

$$\begin{aligned}
K^{-3} \le \|\boldsymbol{a}_i\|_2 \le K^3 \qquad (i \in Q), \\
\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2 \ge K^{-3} \qquad (i, j \in Q,\ i \ne j).
\end{aligned} \tag{83}$$

Similarly, by further calculations deferred to the appendix (Proposition D.7),

$$\varepsilon_{\mathrm{mom}} + \varepsilon_{\mathrm{proj}} + \varepsilon_\perp \le \frac{e^{-s/(2K)}}{512 K^9 \sqrt{n}\, k^5} \le \frac{1}{16 k^2}.$$

All hypotheses of Theorem 6.5 are now satisfied. Applying the theorem gives a non-empty $J \subseteq Q$ and returned matrices of dimension $i_\star = |J|$. There is an orthonormal basis $(\boldsymbol{z}_i)_{i \in J}$ of $\mathbb{R}^{i_\star}$ such that, for every $\boldsymbol{b} \in \mathbb{R}^n$,

$$\left\| \sum_{\ell=1}^n b_\ell \widehat{D}_\ell - \sum_{i \in J} \langle \boldsymbol{b}, \boldsymbol{a}_i \rangle \, \boldsymbol{z}_i \boldsymbol{z}_i^\top \right\|_F \le \varepsilon_{\mathrm{fam}} \, \|\boldsymbol{b}\|_2 , \tag{84}$$

where

$$\varepsilon_{\mathrm{fam}} := 16K^3k^3\left(\varepsilon_{\mathrm{mom}} + \varepsilon_{\mathrm{proj}} + \varepsilon_{\perp}\right).$$

Recovering the normalized regressors. In order to apply Theorem 6.6 when $|J| \geq 2$, we must check the family approximation bound and the norm and separation bounds for the scaled regressors. These are given by (84) and (83), respectively. To obtain the direction accuracy in Theorem 6.1, it remains to bound the resulting error. By the choice of K and the hyperparameter calculation in Proposition D.7,

$$32K^6\sqrt{n}\,k^2\varepsilon_{\mathrm{fam}} \leq e^{-s/(2K)} \leq \frac{\Delta^2}{16B^2}. \tag{85}$$

If $J = \{j\}$, the scalar coordinate matrices give $\widehat{\boldsymbol{a}}_1 = ((\widehat{D}_\ell)_{11})_{\ell\in[n]}$ with $\left\|\widehat{\boldsymbol{a}}_1 - \boldsymbol{a}_j\right\|_2 \leq \varepsilon_{\mathrm{fam}} < K^{-3} \leq \left\|\boldsymbol{a}_j\right\|_2$ by (84), (83), and (85). Normalizing then gives an estimate of the normalized regressor $\boldsymbol{u}_j$ with error at most $2K^3\varepsilon_{\mathrm{fam}}$. For $|J| \geq 2$, applying Theorem 6.6 with failure probability $\delta/2$ gives unit vectors matched bijectively to the normalized regressors indexed by J, with error at most $32K^6\sqrt{n}\,k^2\varepsilon_{\mathrm{fam}}$. Since $\boldsymbol{a}_i/\left\|\boldsymbol{a}_i\right\|_2 = \boldsymbol{w}_i/\left\|\boldsymbol{w}_i\right\|_2$ by Theorem 5.1(iv), both cases give a bijection $\pi : [i_\star] \to J$ satisfying

$$\begin{aligned}
\max_{i\in[i_\star]}\left\|\widehat{\boldsymbol{u}}_i - \frac{\boldsymbol{w}_{\pi(i)}}{\left\|\boldsymbol{w}_{\pi(i)}\right\|_2}\right\|_2 &\leq 32K^6\sqrt{n}\,k^2\varepsilon_{\mathrm{fam}} && \text{(Theorem 6.6 and singleton bound)}\\
&\leq e^{-s/(2K)} \leq \frac{\Delta^2}{16B^2}. && \text{(by (85))}
\end{aligned}$$

By a union bound over all failure events above, the total failure probability is at most δ.

Sample complexity and running time. By Theorem 6.3, EstimateProjectedMoments uses

$$\widetilde{O}\Big((nk)^{O(K^{10})}\log(1/\delta)\Big)$$

samples and runs in time

$$(nk)^{\mathrm{poly}(K)}\log(1/\delta).$$

The observable-operator routines use no further observations. By Theorems 6.5 and 6.6, their total running time is $\mathrm{poly}(n,k)\log(1/\delta)$. Combining these bounds proves the sample and time complexity guarantees. □

7 Restricted Least-Squares

In this section, we show how to refine the estimates of normalized regressors returned by ImplicitTensorDecomposition (Algorithm 6.4) from Section 6 into accurate estimates of the full regressors. Our goal is to prove the result below.

Theorem 7.1 (Restricted Least-Squares Refinement). *Fix $\varepsilon \in (0,1)$, $\delta \in (0,1/3)$, and $i \in [k]$. Suppose the unit vector $\widehat{\boldsymbol{u}}$ satisfies*

$$\left\|\widehat{\boldsymbol{u}} - \frac{\boldsymbol{w}_i}{\left\|\boldsymbol{w}_i\right\|_2}\right\|_2 \leq \frac{\Delta^2}{16B^2}.$$

Then RESTRICTEDLEASTSQUARES *(Algorithm 7.1) returns an estimate* $\widehat{w}$ *satisfying*

$$\|\widehat{w} - w_i\|_2 \le \varepsilon \tag{86}$$

with probability at least $1-\delta$*. Its sample complexity is*

$$\widetilde{O}\left(\left(\frac{nk}{\varepsilon}\right)^{5K^2}\log(1/\delta)\right)$$

and its running time is

$$\widetilde{O}\left(\left(\frac{nk}{\varepsilon}\right)^{\mathrm{poly}(K)}\log(1/\delta)\right).$$

Let $\mathcal{E}_{\widehat{u}} = \{\Gamma \le \langle \widehat{u}, X\rangle \le \Gamma + 1\}$ be the large-projection event, with $\Gamma = \Theta(\sqrt{\log(n/\varepsilon)})$ chosen as in Algorithm 7.1. When $\widehat{u}$ is close to $w_i/\|w_i\|_2$, the response Y_i is likely to attain the maximum on this event. We therefore estimate the least-squares solution restricted to this event. Its normal equations are $\Sigma_{\widehat{u}} w = b_{\widehat{u}}$, where

$$\begin{aligned} \Sigma_{\widehat{u}} &:= \mathbb{E}_X\left[XX^\top \mathbf{1}_{\mathcal{E}_{\widehat{u}}}\right], \\ b_{\widehat{u}} &:= \mathbb{E}_{X,\xi}\left[XY_{\max}\mathbf{1}_{\mathcal{E}_{\widehat{u}}}\right]. \end{aligned} \tag{87}$$

Our algorithm computes $\Sigma_{\widehat{u}}^{-1}$ in closed form and applies it to a trimmed-mean estimate of $b_{\widehat{u}}$. Its pseudocode can be found in Algorithm 7.1.

Algorithm 7.1: RESTRICTEDLEASTSQUARES

Input: A unit vector $\widehat{u}$, accuracy $\varepsilon \in (0,1)$, and failure probability $\delta \in (0, 1/3)$
Output: An estimate $\widehat{w} \in \mathbb{R}^n$

$\Gamma \leftarrow 2K\sqrt{\log(\frac{n}{\varepsilon})}$;
$\Pr[\mathcal{E}_{\widehat{u}}] \leftarrow \Phi(\Gamma+1) - \Phi(\Gamma)$;
$\nu_2 \leftarrow \Pr[\mathcal{E}_{\widehat{u}}] + \Gamma\phi(\Gamma) - (\Gamma+1)\phi(\Gamma+1)$;
$\Sigma_{\widehat{u}}^{-1} \leftarrow \nu_2^{-1}\widehat{u}\widehat{u}^\top + \frac{1}{\Pr[\mathcal{E}_{\widehat{u}}]}(I - \widehat{u}\widehat{u}^\top)$;
$\mathrm{v}_{\mathrm{ref}} \leftarrow 96B^2\log k$; `// coordinate second-moment bound`
`// Lemma 4.7 and Appendix B.2`
$\widehat{b}_{\widehat{u}} \leftarrow \mathtt{TrimmedMean}(XY_{\max}\mathbf{1}\{\Gamma \le \langle \widehat{u}, X\rangle \le \Gamma+1\}; \mathrm{v}_{\mathrm{ref}}, \frac{\Pr[\mathcal{E}_{\widehat{u}}]\varepsilon}{2\sqrt{n}}, \delta)$;
$\widehat{w} \leftarrow \Sigma_{\widehat{u}}^{-1}\widehat{b}_{\widehat{u}}$;
return $\widehat{w}$;

The rest of this section is dedicated to proving Theorem 7.1.

7.1 Proof of Theorem 7.1

Before analyzing Algorithm 7.1, we bound the conditional second moments of $\langle v, X\rangle$ and $Y_{\max} - Y_i$ to control the bias from using the maximum response. We also bound the second moment of each coordinate of $XY_{\max}\mathbf{1}_{\mathcal{E}_{\widehat{u}}}$ to control the sampling error.

Lemma 7.2 (Moment Bounds for Restricted Least Squares)**.** *Consider the Gaussian self-selection model (Definition 1.1) under the boundedness assumption (Assumption 1.2). Use K from Hyperparameters 4.1, and let $\Gamma \geq 1$. Fix a unit vector $\widehat{\boldsymbol{u}}$ and $i \in [k]$ such that*

$$\langle \boldsymbol{w}_i - \boldsymbol{w}_j, \widehat{\boldsymbol{u}} \rangle \geq \frac{\Delta^2}{2B} \qquad (j \neq i).$$

Then for every unit vector $\boldsymbol{v}$ and every $j \in [n]$,

$$\begin{aligned} \mathbb{E}_{\boldsymbol{X}}[\langle \boldsymbol{v}, \boldsymbol{X} \rangle^2 \mid \mathcal{E}_{\widehat{\boldsymbol{u}}}] &\leq 4\Gamma^2\,, \\ \mathbb{E}_{\boldsymbol{X},\xi}[(Y_{\max} - Y_i)^2 \mid \mathcal{E}_{\widehat{\boldsymbol{u}}}] &\leq 6B^2 k e^{-2\Gamma^2/K}\,, \\ \mathbb{E}_{\boldsymbol{X},\xi}[(X_j Y_{\max} \mathbf{1}_{\mathcal{E}_{\widehat{\boldsymbol{u}}}})^2] &\leq 96B^2 \log k\,. \end{aligned}$$

Proof. We first bound the conditional covariate and response moments, then bound the coordinate second moments used by TRIMMEDMEAN. By the choice of Hyperparameters 4.1,

$$K\Delta^4 \geq 16(1+B^2)^3 \geq 16B^2(4B^2+2). \tag{88}$$

The last inequality uses $(1+B^2)^3 - B^2(4B^2+2) = B^4(B^2-1) + B^2 + 1 \geq 0$.

Conditional covariate moment. Fix a unit vector $\boldsymbol{v}$ and $y \in [\Gamma, \Gamma+1]$. The conditional Gaussian formula (Proposition 4.3) gives

$$\begin{aligned} \mathbb{E}_{\boldsymbol{X}}[\langle \boldsymbol{v}, \boldsymbol{X} \rangle^2 \mid \langle \widehat{\boldsymbol{u}}, \boldsymbol{X} \rangle = y] &= y^2 \langle \boldsymbol{v}, \widehat{\boldsymbol{u}} \rangle^2 + \left\| (I - \widehat{\boldsymbol{u}}\widehat{\boldsymbol{u}}^\top)\boldsymbol{v} \right\|_2^2 \\ &= 1 + (y^2 - 1)\langle \boldsymbol{v}, \widehat{\boldsymbol{u}} \rangle^2 && (\|\boldsymbol{v}\|_2 = 1) \\ &\leq y^2 && (y \geq 1;\ |\langle \boldsymbol{v}, \widehat{\boldsymbol{u}} \rangle| \leq 1) \\ &\leq (\Gamma+1)^2 \leq 4\Gamma^2. && (y \leq \Gamma+1;\ \Gamma \geq 1) \end{aligned}$$

Averaging over $y \in [\Gamma, \Gamma+1]$ conditional on $\mathcal{E}_{\widehat{\boldsymbol{u}}}$ proves the first bound.

Conditional response difference. The truncated Gaussian second-moment bound (Proposition A.3) gives

$$\mathbb{E}_G[G^2 \mathbf{1}_{\{G>0\}}] \leq \mathrm{Var}_G(G)\mathbb{P}_G(G > 0)$$

for every nondegenerate Gaussian G with $\mathbb{E}_G[G] \leq 0$. We apply this bound to the conditional response differences.

Fix $j \neq i$ and $y \in [\Gamma, \Gamma+1]$. Applying the conditional Gaussian formula (Proposition 4.3) to

$$Y_j - Y_i = \langle \boldsymbol{w}_j - \boldsymbol{w}_i, \boldsymbol{X} \rangle + \xi_j - \xi_i$$

gives

$$Y_j - Y_i \mid \langle \widehat{\boldsymbol{u}}, \boldsymbol{X} \rangle = y \sim N\left(y \langle \boldsymbol{w}_j - \boldsymbol{w}_i, \widehat{\boldsymbol{u}} \rangle\,, \left\| (I - \widehat{\boldsymbol{u}}\widehat{\boldsymbol{u}}^\top)(\boldsymbol{w}_j - \boldsymbol{w}_i) \right\|_2^2 + 2 \right). \tag{89}$$

Here the noises are independent of $\boldsymbol{X}$, and $\mathrm{Var}(\xi_j - \xi_i) = 1 + 1 = 2$. The conditional variance is therefore positive. The assumed directional gap and boundedness give

$$\mathbb{E}_{\boldsymbol{X},\xi}[Y_j - Y_i \mid \langle \widehat{\boldsymbol{u}}, \boldsymbol{X} \rangle = y] = y \langle \boldsymbol{w}_j - \boldsymbol{w}_i, \widehat{\boldsymbol{u}} \rangle \leq -\frac{y\Delta^2}{2B}, \tag{90}$$

$$\begin{aligned} \mathrm{Var}_{\boldsymbol{X},\xi}(Y_j - Y_i \mid \langle \widehat{\boldsymbol{u}}, \boldsymbol{X} \rangle = y) &\leq \|\boldsymbol{w}_j - \boldsymbol{w}_i\|_2^2 + 2 \\ &\leq 4B^2 + 2. \end{aligned} \tag{91}$$

Since $Y_{\max} = \max_{j\in[k]} Y_j$,

$$(Y_{\max} - Y_i)^2 \le \sum_{j\neq i} (Y_j - Y_i)^2 \mathbf{1}_{\{Y_j > Y_i\}}.$$

The truncated Gaussian second-moment bound and a Gaussian tail bound (Proposition 4.2) therefore give

$$\begin{aligned}
&\mathbb{E}_{X,\xi}[(Y_{\max} - Y_i)^2 \mid \langle \widehat{u}, X\rangle = y] \\
&\quad\le \sum_{j\neq i} \mathbb{E}_{X,\xi}\left[(Y_j - Y_i)^2 \mathbf{1}_{\{Y_j > Y_i\}} \,\middle|\, \langle \widehat{u}, X\rangle = y\right] \\
&\quad\le \sum_{j\neq i} \mathrm{Var}_{X,\xi}(Y_j - Y_i \mid \langle \widehat{u}, X\rangle = y)\, \mathbb{P}_{X,\xi}(Y_j > Y_i \mid \langle \widehat{u}, X\rangle = y) && \text{(by Proposition A.3)} \\
&\quad\le (4B^2+2) \sum_{j\neq i} \mathbb{P}_{X,\xi}(Y_j > Y_i \mid \langle \widehat{u}, X\rangle = y) && \text{(by (91))} \\
&\quad\le (4B^2+2) k \exp\left(-\frac{y^2\Delta^4}{8B^2(4B^2+2)}\right) && \text{(Gaussian tail bound; by (90), (91))} \\
&\quad\le (4B^2+2) k e^{-2y^2/K} && \text{(choice of } K \text{ (88))} \\
&\quad\le 6B^2 k e^{-2\Gamma^2/K}. && (B \ge 1;\, y \ge \Gamma)
\end{aligned}$$

Averaging over $y \in [\Gamma, \Gamma+1]$ conditional on $\mathcal{E}_{\widehat{u}}$ proves the second bound.

Coordinate second moment. It remains to control the second moment of $X_j Y_{\max} \mathbf{1}_{\mathcal{E}_{\widehat{u}}}$ used by TRIMMEDMEAN. The coordinate X_j is standard Gaussian, and the responses Y_i are centered Gaussian. Since $|Y_{\max}| \le \max_{i\in[k]} |Y_i|$, Cauchy–Schwarz and the Gaussian maximum bound (Proposition A.2) give

$$\begin{aligned}
&\mathbb{E}_{X,\xi}[(X_j Y_{\max} \mathbf{1}_{\mathcal{E}_{\widehat{u}}})^2] \\
&\quad\le \sqrt{\mathbb{E}_X[X_j^4]} \left(\mathbb{E}_{X,\xi} \max_{i\in[k]} |Y_i|^4\right)^{1/2} && \text{(Cauchy–Schwarz)} \\
&\quad= \sqrt{3} \left(\mathbb{E}_{X,\xi} \max_{i\in[k]} |Y_i|^4\right)^{1/2} && \text{(standard Gaussian fourth moment)} \\
&\quad\le \sqrt{3} \left(3\sqrt{\log k} \left(\max_{i\in[k]} \mathrm{Var}_{X,\xi}(Y_i)\right)^{1/2}\right)^2 && \text{(by Proposition A.2)} \\
&\quad= 9\sqrt{3} \log k \max_{i\in[k]} (1 + \|w_i\|_2^2) \\
&\quad\le 18\sqrt{3} B^2 \log k && \text{(boundedness (Assumption 1.2); } B \ge 1) \\
&\quad\le 96 B^2 \log k.
\end{aligned}$$

This proves the third bound. □

We are now ready to prove the guarantees of our RESTRICTEDLEASTSQUARES algorithm.

Proof of Theorem 7.1. We first compute the inverse of the restricted second moment and obtain a bound on its operator norm. This controls blowup in the bias from replacing Y_i by $Y_{\max}$ and the sampling error in the restricted response when solving the normal equations.

Restricted second moment. To compute the second restricted moment in closed form, we first evaluate the following scalar quantities:

$$\begin{aligned}
\Pr[\mathcal{E}_{\widehat{u}}] &= \Phi(\Gamma+1) - \Phi(\Gamma), \\
\nu_2 &:= \int_{\Gamma}^{\Gamma+1} y^2 \phi(y)\,\mathrm{d}y \\
&= -\int_{\Gamma}^{\Gamma+1} y\phi'(y)\,\mathrm{d}y && (\phi'(y) = -y\phi(y)) \\
&= -[y\phi(y)]_{\Gamma}^{\Gamma+1} + \int_{\Gamma}^{\Gamma+1} \phi(y)\,\mathrm{d}y && \text{(integration by parts)} \\
&= \Pr[\mathcal{E}_{\widehat{u}}] + \Gamma\phi(\Gamma) - (\Gamma+1)\phi(\Gamma+1). && (92)
\end{aligned}$$

The conditional Gaussian formula (Proposition 4.3) gives mean $y\widehat{u}$ and covariance $I - \widehat{u}\widehat{u}^\top$ conditional on $\langle \widehat{u}, \boldsymbol{X}\rangle = y$. The second-moment formula for a random vector therefore gives

$$\begin{aligned}
\Sigma_{\widehat{u}} &= \int_{\Gamma}^{\Gamma+1} \mathbb{E}_{\boldsymbol{X}}[\boldsymbol{X}\boldsymbol{X}^\top \mid \langle \widehat{u}, \boldsymbol{X}\rangle = y]\phi(y)\,\mathrm{d}y \\
&= \int_{\Gamma}^{\Gamma+1} (y^2\widehat{u}\widehat{u}^\top + I - \widehat{u}\widehat{u}^\top)\phi(y)\,\mathrm{d}y && \text{(second-moment formula)} \\
&= \nu_2\widehat{u}\widehat{u}^\top + \Pr[\mathcal{E}_{\widehat{u}}](I - \widehat{u}\widehat{u}^\top). && (93)
\end{aligned}$$

This matrix acts by ν_2 along $\widehat{u}$ and by $\Pr[\mathcal{E}_{\widehat{u}}]$ on its orthogonal complement. The choice of Γ gives $\Gamma \geq \max\{1, \sqrt{\log k}\}$, so $\nu_2 \geq \Pr[\mathcal{E}_{\widehat{u}}] > 0$. Hence

$$\begin{aligned}
\Sigma_{\widehat{u}}^{-1} &= \nu_2^{-1}\widehat{u}\widehat{u}^\top + \frac{1}{\Pr[\mathcal{E}_{\widehat{u}}]}(I - \widehat{u}\widehat{u}^\top), \\
\left\|\Sigma_{\widehat{u}}^{-1}\right\|_{\mathrm{op}} &\leq \frac{1}{\Pr[\mathcal{E}_{\widehat{u}}]}.
\end{aligned} \tag{94}$$

Error decomposition. Since the noise ξ_i is centered and independent of $\boldsymbol{X}$,

$$\mathbb{E}_{\boldsymbol{X},\xi_i}[\boldsymbol{X}Y_i\mathbf{1}_{\mathcal{E}_{\widehat{u}}}] = \mathbb{E}_{\boldsymbol{X}}[\boldsymbol{X}\boldsymbol{X}^\top\mathbf{1}_{\mathcal{E}_{\widehat{u}}}]w_i = \Sigma_{\widehat{u}}w_i.$$

Thus the bias in the restricted response is

$$\boldsymbol{e}_{\widehat{u}} := \boldsymbol{b}_{\widehat{u}} - \Sigma_{\widehat{u}}w_i = \mathbb{E}_{\boldsymbol{X},\xi}\left[\boldsymbol{X}(Y_{\max} - Y_i)\mathbf{1}_{\mathcal{E}_{\widehat{u}}}\right]. \tag{95}$$

The integrand vanishes whenever $Y_{\max} = Y_i$. Since $\widehat{w} = \Sigma_{\widehat{u}}^{-1}\widehat{\boldsymbol{b}}_{\widehat{u}}$, adding and subtracting $\boldsymbol{b}_{\widehat{u}}$ gives

$$\begin{aligned}
\|\widehat{w} - w_i\|_2 &= \left\|\Sigma_{\widehat{u}}^{-1}\left(\boldsymbol{e}_{\widehat{u}} + \widehat{\boldsymbol{b}}_{\widehat{u}} - \boldsymbol{b}_{\widehat{u}}\right)\right\|_2 \\
&\leq \left\|\Sigma_{\widehat{u}}^{-1}\right\|_{\mathrm{op}} \left\|\boldsymbol{e}_{\widehat{u}} + \widehat{\boldsymbol{b}}_{\widehat{u}} - \boldsymbol{b}_{\widehat{u}}\right\|_2 \\
&\leq \left\|\Sigma_{\widehat{u}}^{-1}\right\|_{\mathrm{op}} \left(\|\boldsymbol{e}_{\widehat{u}}\|_2 + \left\|\widehat{\boldsymbol{b}}_{\widehat{u}} - \boldsymbol{b}_{\widehat{u}}\right\|_2\right) && \text{(triangle inequality)} \\
&\leq \underbrace{\frac{\|\boldsymbol{e}_{\widehat{u}}\|_2}{\Pr[\mathcal{E}_{\widehat{u}}]}}_{\text{bias}} + \underbrace{\frac{\left\|\widehat{\boldsymbol{b}}_{\widehat{u}} - \boldsymbol{b}_{\widehat{u}}\right\|_2}{\Pr[\mathcal{E}_{\widehat{u}}]}}_{\text{sampling error}}. && \text{(by (94))} \quad (96)
\end{aligned}$$

We will show that each term is at most $\varepsilon/2$.

Conditional mean gap. We first verify the hypothesis of Lemma 7.2. Recall $\boldsymbol{u}_i = \boldsymbol{w}_i / \|\boldsymbol{w}_i\|_2$. Boundedness and uncoveredness (Assumption 1.2) give

$$\begin{aligned}
\langle \boldsymbol{w}_i - \boldsymbol{w}_j, \widehat{\boldsymbol{u}} \rangle &= \langle \boldsymbol{w}_i - \boldsymbol{w}_j, \boldsymbol{u}_i \rangle + \langle \boldsymbol{w}_i - \boldsymbol{w}_j, \widehat{\boldsymbol{u}} - \boldsymbol{u}_i \rangle \\
&\geq \langle \boldsymbol{w}_i - \boldsymbol{w}_j, \boldsymbol{u}_i \rangle - \|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2 \, \|\widehat{\boldsymbol{u}} - \boldsymbol{u}_i\|_2 && \text{(Cauchy–Schwarz)} \\
&\geq \frac{\Delta^2}{B} - 2B\frac{\Delta^2}{16B^2} && \text{(by Assumption 1.2; estimation assumption)} \\
&\geq \frac{\Delta^2}{2B} \qquad (j \neq i). && (97)
\end{aligned}$$

Bias. We apply Cauchy–Schwarz to the conditional expectation. Having verified the mean-gap hypothesis, we will apply Lemma 7.2 to bound the resulting second moments. For every unit vector $\boldsymbol{v}$,

$$\begin{aligned}
\frac{|\langle \boldsymbol{v}, \boldsymbol{e}_{\widehat{\boldsymbol{u}}} \rangle|}{\Pr[\mathcal{E}_{\widehat{\boldsymbol{u}}}]} &= \frac{1}{\Pr[\mathcal{E}_{\widehat{\boldsymbol{u}}}]} \left| \mathbb{E}_{\boldsymbol{X},\xi} \left[\langle \boldsymbol{v}, \boldsymbol{X} \rangle \, (Y_{\max} - Y_i) \mathbf{1}_{\mathcal{E}_{\widehat{\boldsymbol{u}}}} \right] \right| && \text{(definition of } \boldsymbol{e}_{\widehat{\boldsymbol{u}}}\text{; linearity)} \\
&= \left| \mathbb{E}_{\boldsymbol{X},\xi} \left[\langle \boldsymbol{v}, \boldsymbol{X} \rangle \, (Y_{\max} - Y_i) \mid \mathcal{E}_{\widehat{\boldsymbol{u}}} \right] \right| && \text{(conditional expectation)} \\
&\leq \left(\mathbb{E}_{\boldsymbol{X}} [\langle \boldsymbol{v}, \boldsymbol{X} \rangle^2 \mid \mathcal{E}_{\widehat{\boldsymbol{u}}}] \right)^{1/2} \left(\mathbb{E}_{\boldsymbol{X},\xi} [(Y_{\max} - Y_i)^2 \mid \mathcal{E}_{\widehat{\boldsymbol{u}}}] \right)^{1/2} && \text{(Cauchy–Schwarz)} \\
&\leq \sqrt{4\Gamma^2} \sqrt{6B^2 k e^{-2\Gamma^2/K}} && \text{(by Lemma 7.2)} \\
&= 2\sqrt{6}\, B\Gamma \sqrt{k}\, e^{-\Gamma^2/K}.
\end{aligned}$$

Taking the supremum over unit vectors gives

$$\begin{aligned}
\frac{\|\boldsymbol{e}_{\widehat{\boldsymbol{u}}}\|_2}{\Pr[\mathcal{E}_{\widehat{\boldsymbol{u}}}]} &\leq 2\sqrt{6}\, B\Gamma \sqrt{k}\, e^{-\Gamma^2/K} \\
&\leq 5B\Gamma \sqrt{k}\, e^{-\Gamma^2/K} && (2\sqrt{6} \leq 5) \\
&\leq \frac{\varepsilon}{2}. && \text{(by Proposition E.1)} \quad (98)
\end{aligned}$$

Sampling error. The bias is now at most $\varepsilon/2$. To control the sampling term in (96), we apply the trimmed-mean guarantee (Lemma 4.7) to $\boldsymbol{X} Y_{\max} \mathbf{1}_{\mathcal{E}_{\widehat{\boldsymbol{u}}}}$ with coordinate accuracy $\Pr[\mathcal{E}_{\widehat{\boldsymbol{u}}}]\varepsilon/(2\sqrt{n})$. This requires a second-moment bound for every coordinate. By Lemma 7.2,

$$\mathbb{E}_{\boldsymbol{X},\xi}[(X_j Y_{\max} \mathbf{1}_{\mathcal{E}_{\widehat{\boldsymbol{u}}}})^2] \leq 96B^2 \log k = \mathsf{v}_{\mathrm{ref}} \qquad (j \in [n]).$$

The requested accuracy is admissible because

$$0 < \frac{\Pr[\mathcal{E}_{\widehat{\boldsymbol{u}}}]\varepsilon}{2\sqrt{n}} \leq \frac{1}{2} \leq \sqrt{\mathsf{v}_{\mathrm{ref}}}.$$

Thus, with probability at least $1 - \delta$,

$$\begin{aligned}
\left\| \widehat{\boldsymbol{b}}_{\widehat{\boldsymbol{u}}} - \boldsymbol{b}_{\widehat{\boldsymbol{u}}} \right\|_2 &\leq \sqrt{n}\, \frac{\Pr[\mathcal{E}_{\widehat{\boldsymbol{u}}}]\varepsilon}{2\sqrt{n}} = \frac{\Pr[\mathcal{E}_{\widehat{\boldsymbol{u}}}]\varepsilon}{2}, \\
\frac{\left\| \widehat{\boldsymbol{b}}_{\widehat{\boldsymbol{u}}} - \boldsymbol{b}_{\widehat{\boldsymbol{u}}} \right\|_2}{\Pr[\mathcal{E}_{\widehat{\boldsymbol{u}}}]} &\leq \frac{\varepsilon}{2}. && (99)
\end{aligned}$$

Combining these bounds in (96) proves $\|\widehat{w} - w_i\|_2 \le \varepsilon$.

Sample complexity and running time. To bound the sample complexity, we need a lower bound on $\Pr[\mathcal{E}_{\widehat{u}}]$. We integrate over $[\Gamma, \Gamma + \Gamma^{-1}]$, where the Gaussian density stays within a constant factor of $\phi(\Gamma)$. For $0 \le h \le \Gamma^{-1}$, the bound $\Gamma \ge 1$ gives

$$\phi(\Gamma + h) = \phi(\Gamma) e^{-\Gamma h - h^2/2} \ge e^{-3/2} \phi(\Gamma).$$

Consequently,

$$\begin{aligned}
\Pr[\mathcal{E}_{\widehat{u}}] &= \int_{\Gamma}^{\Gamma+1} \phi(y)\, \mathrm{d}y \\
&\ge \int_{0}^{\Gamma^{-1}} \phi(\Gamma + h)\, \mathrm{d}h && (\Gamma^{-1} \le 1) \\
&\ge \frac{e^{-3/2}\phi(\Gamma)}{\Gamma} \\
&\ge \frac{e^{-\Gamma^2/2}}{16\Gamma}. && (e^{3/2}\sqrt{2\pi} \le 16) \quad (100)
\end{aligned}$$

Substituting (100) into the trimmed-mean sample bound (Lemma 4.7) gives

$$O\left(\frac{nB^2 \log k}{\varepsilon^2} \Gamma^2 e^{\Gamma^2} \log\left(\frac{n}{\delta}\right) \right)$$

observations. By the choice of K and Γ and the hyperparameter calculations deferred to the appendix (Proposition E.1), the trimmed-mean estimate uses

$$\widetilde{O}\left(\left(\frac{nk}{\varepsilon}\right)^{5K^2} \log(1/\delta) \right)$$

observations. All operations run in polynomial time with respect to the number of samples so the algorithm terminates in

$$\widetilde{O}\left(\left(\frac{nk}{\varepsilon}\right)^{\mathrm{poly}(K)} \log(1/\delta) \right)$$

time. □

8 Moment Peeling & Estimation

A key subroutine of our implicit tensor decomposition algorithm from Section 6 is a mean estimation oracle for the projected Hermite moments of the unrecovered regressors. For a set Q of unrecovered regressors, recall that $Y_{\max,Q} = \max_{i \in Q} Y_i$ and that its degree-d restricted Hermite moment is

$$\mathsf{T}_d^Q = \mathbb{E}_{X,\xi_Q}[\mathsf{H}_d(X)\mathbf{1}\{T \le Y_{\max,Q} \le T+1\}].$$

In this section, we implement this oracle given samples of the form $(X, U\widetilde{X}^{\otimes r}, Y_{\max})$ where U is some projection matrix. Our goal is to prove the following result, which quantifies the estimation error relative to the trace $\mathrm{tr}(A_Q^{(r)}) = \sum_{i \in Q} \lambda_i^{(2r)}$.

Theorem 8.1 (Hermite Moment Estimation by Peeling)**.** *Consider the Gaussian self-selection model (Definition 1.1) under the boundedness and uncoveredness assumptions (Assumption 1.2). Use the hyperparameters K, T, s from Hyperparameters 4.1.*

Let $2 \le r \le s$, and fix a projection map $U : \mathbb{R}^{n^r} \to \mathbb{R}^p$, where $U \in \mathbb{R}^{p \times n^r}$ has orthonormal rows. Assume we are given fresh samples $(\boldsymbol{X}, U\widetilde{\boldsymbol{X}}^{\otimes r}, Y_{\max})$ where $\widetilde{\boldsymbol{X}} = \boldsymbol{X} + \mathrm{i}\boldsymbol{X}'$.

Fix a sampling accuracy $\varepsilon_{\mathrm{est}} \in (0,1)$ and a failure probability $\delta \in (0, 1/3)$. Let $\widehat{\mathcal{R}}$ be a set of regressor estimates matched to some $R \subsetneq [k]$, and put $Q = [k] \setminus R$. If the estimates satisfy

$$\left\| \widehat{\boldsymbol{w}} - \boldsymbol{w}_{\mathrm{match}(\widehat{\boldsymbol{w}})} \right\|_2 \le e^{-K^6 s} \qquad (\widehat{\boldsymbol{w}} \in \widehat{\mathcal{R}}),$$

then the following guarantees hold.

(i) For the projected even moment at order r, EstimateMomentsByPeeling *(Algorithm 8.1), given sampling accuracy $\varepsilon_{\mathrm{est}}$ and confidence parameter δ, returns a symmetric matrix $\widehat{M} \in \mathbb{R}^{p \times p}$ satisfying, with probability at least $1 - \delta$,*

$$\left\| \widehat{M} - U \,\mathrm{Mat}_r(\mathsf{T}^Q_{2r}) U^\top \right\|_F \le \varepsilon_{\mathrm{est}} + \frac{\sqrt{k}}{\sqrt{2\pi}} e^{-K^6 s} + e^{4s \log K - T^2/(2K)} k \,\mathrm{tr}(A^{(r)}_Q). \tag{101}$$

(ii) For the projected odd moment at $r = s$, EstimateMomentsByPeeling *returns symmetric matrices $\widehat{M}_1, \dots, \widehat{M}_n \in \mathbb{R}^{p \times p}$ satisfying, with probability at least $1 - \delta$, simultaneously for every $\boldsymbol{b} \in \mathbb{R}^n$,*

$$\left\| \sum_{\ell=1}^n b_\ell \widehat{M}_\ell - U \,\mathrm{Mat}_s(\mathsf{T}^Q_{2s+1} \lrcorner \boldsymbol{b}) U^\top \right\|_F \le \left(\varepsilon_{\mathrm{est}} + \frac{\sqrt{k}}{\sqrt{2\pi}} e^{-K^6 s} + e^{4s \log K - T^2/(2K)} k \,\mathrm{tr}(A^{(s)}_Q) \right) \|\boldsymbol{b}\|_2 . \tag{102}$$

Each call requires at most

$$\widetilde{O}\left(\frac{p^2 n}{\varepsilon_{\mathrm{est}}^2} 100^s \log(1/\delta) \right)$$

projected samples. Moreover, its running time is at most $\mathrm{poly}(n, k, p, 1/\varepsilon_{\mathrm{est}}, \log(1/\delta))$.

As we will see in the roadmap (Section 8.1), the estimation error can be decomposed into three terms. The rest of this section aims to bound each of these terms which can then be combined into a proof of Theorem 8.1.

8.1 Proof Roadmap

Our estimation algorithm is easiest to present and analyze for expectations of the form

$$\mathbb{E}_{\boldsymbol{X}, \xi, \boldsymbol{X}'}[\Psi(\boldsymbol{X}, \boldsymbol{X}') \mathbf{1}\{T \le Y_{\max, Q} \le T + 1\}],$$

where $\Psi \in \mathbb{R}^{p \times p}$ is a random matrix. We will then apply our analysis with Ψ being projections of the even and odd Hermite moments.

One key challenge is that we do not directly observe $Y_{\max,Q}$ and hence we must use a proxy $\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X}, Y_{\max})$ for the indicator variable $\mathbf{1}\{T \le Y_{\max,Q} \le T+1\}$. We define this proxy in Section 8.1.1. Then, we discuss the specific instantiations of Ψ in Section 8.1.2. We then decompose the estimation error into three terms in Section 8.1.3 and state our bounds for each term.

8.1.1 Signed Peeling Weight

We now define the proxy of the indicator $\mathbf{1}\{T \le Y_{\max,Q} \le T+1\}$. Let $\widehat{\mathcal{R}}$ be the list of recovered regressors corresponding to indices $R \subsetneq [k]$ and put $Q = [k] \setminus R$.

Ideal Peeling Weight. We first define an idealized peeling weight $\omega_R(\boldsymbol{X}, Y_{\max})$ assuming we know the exact regressors $\boldsymbol{w}_i, i \in R$. Recall for an index set $I \subseteq [k]$, we write $Y_{\max,I} = \max_{i\in I} Y_i$ to denote its maximum response. Define the conditional CDF of the maximum response conditioned on the covariate, which can be written as a product of CDFs of the standard Gaussian

$$F_I(t \mid \boldsymbol{x}) := \mathbb{P}_{\xi_I}(Y_{\max,I} \le t \mid \boldsymbol{X} = \boldsymbol{x}) = \prod_{i\in I} \Phi\Big(t - \boldsymbol{w}_i^\top \boldsymbol{x}\Big) \,. \tag{103}$$

Recall the peeling identity from Section 2.2.5. Since $Y_{\max} = \max\{Y_{\max,R}, Y_{\max,Q}\}$,

$$\begin{aligned}\mathbf{1}\{T \le Y_{\max} \le T+1\} - \mathbf{1}\{T \le Y_{\max,R} \le T+1\} &= \mathbf{1}\{T \le Y_{\max,Q} \le T+1\} \\ &\quad - \mathbf{1}\{T \le \min\{Y_{\max,R}, Y_{\max,Q}\} \le T+1\}.\end{aligned}$$

We would like to define the proxy $\omega_R(\boldsymbol{X}, Y_{\max})$ to be the LHS, since it is equal to the indicator we want and a term we will show to be negligible. However, we do not directly observe $Y_{\max,R}$. Our saving grace is that we only need equality under expectation so that we can define the proxy to match the LHS in conditional expectation conditioned on the covariate $\boldsymbol{X}$.

$$\begin{aligned}\omega_R(\boldsymbol{X}, Y_{\max}) &:= \mathbf{1}\{T \le Y_{\max} \le T+1\} - \big(F_R(T+1 \mid \boldsymbol{X}) - F_R(T \mid \boldsymbol{X})\big)\,, \\ \mathbb{E}_\xi[\omega_R(\boldsymbol{X}, Y_{\max}) \mid \boldsymbol{X}] &= \mathbb{E}_\xi[\mathbf{1}\{T \le Y_{\max} \le T+1\} - \mathbf{1}\{T \le Y_{\max,R} \le T+1\} \mid \boldsymbol{X}]\,.\end{aligned} \tag{104}$$

Peeling Weight. In reality, we only have a list $\widehat{\mathcal{R}}$ of estimates to the exact regressors $\boldsymbol{w}_i, i \in R$. Since we only need said regressors to evaluate the conditional CDF, we can simply plug in our estimates instead. Define the conditional CDF

$$\widehat{F}_{\widehat{\mathcal{R}}}(t \mid \boldsymbol{x}) := \prod_{\widehat{\boldsymbol{w}}\in\widehat{\mathcal{R}}} \Phi(t - \langle \widehat{\boldsymbol{w}}, \boldsymbol{x}\rangle) \,. \tag{105}$$

We define the signed peeling weight to be

$$\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X}, Y_{\max}) := \mathbf{1}\{T \le Y_{\max} \le T+1\} - \big(\widehat{F}_{\widehat{\mathcal{R}}}(T+1 \mid \boldsymbol{X}) - \widehat{F}_{\widehat{\mathcal{R}}}(T \mid \boldsymbol{X})\big) \in [-1, 1]\,. \tag{106}$$

8.1.2 Projected Hermite Features

We now discuss the specific projected Hermite features Ψ whose restricted expectations we wish to estimate. For $1 \le r \le s$, fix a projection map $U : \mathbb{R}^{n^r} \to \mathbb{R}^p$ with orthonormal rows, so $UU^\top = I_p$. Recall that part of our implicit tensor decomposition algorithm presented in Section 6.2 is an iterative projection procedure that recursively computes the projected moments and projection maps degree by degree. Extending the construction by one tensor degree requires an even moment estimate to determine the next selected subspace.

Define the functions used by EstimateMomentsByPeeling as

$$\Psi^U_{\text{even}}(\boldsymbol{X}, \boldsymbol{X}') := \frac{\operatorname{Re}\Big[(U\widetilde{\boldsymbol{X}}^{\otimes r})(U\widetilde{\boldsymbol{X}}^{\otimes r})^\top\Big]}{\sqrt{(2r)!}} \tag{107}$$

and

$$\Psi^U_{\mathrm{odd},\ell}(\boldsymbol{X},\boldsymbol{X}') := \frac{\mathrm{Re}\Big[(U\widetilde{\boldsymbol{X}}^{\otimes r})(U\widetilde{\boldsymbol{X}}^{\otimes r})^\top \widetilde{X}_\ell\Big]}{\sqrt{(2r+1)!}}, \qquad \ell \in [n]. \tag{108}$$

These definitions apply to every $r \geq 1$ and give real symmetric $p \times p$ matrices. Here $\widetilde{\boldsymbol{X}} = \boldsymbol{X} + \mathrm{i}\boldsymbol{X}'$ denotes the complexified covariate. Averaging over the complex component $\boldsymbol{X}'$ gives

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}'}[\Psi^U_{\mathrm{even}}(\boldsymbol{X},\boldsymbol{X}') \mid \boldsymbol{X}] &= U\,\mathrm{Mat}_r(\mathsf{H}_{2r}(\boldsymbol{X}))U^\top, \\
\mathbb{E}_{\boldsymbol{X}'}[\Psi^U_{\mathrm{odd},\ell}(\boldsymbol{X},\boldsymbol{X}') \mid \boldsymbol{X}] &= U\,\mathrm{Mat}_r(\mathsf{H}_{2r+1}(\boldsymbol{X}) \lrcorner \boldsymbol{e}_\ell)U^\top.
\end{aligned} \tag{109}$$

Our algorithm estimates the weighted expectations below. In this display, the projected Hermite features are evaluated at $(\boldsymbol{X},\boldsymbol{X}')$.

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X},\xi,\boldsymbol{X}'}\Big[\Psi^U_{\mathrm{even}}\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X},Y_{\max})\Big] &\approx \mathbb{E}_{\boldsymbol{X},\xi_Q,\boldsymbol{X}'}\Big[\Psi^U_{\mathrm{even}}\mathbf{1}\{T \leq Y_{\max,Q} \leq T+1\}\Big] \\
&= U\,\mathrm{Mat}_r(\mathsf{T}^Q_{2r})U^\top, \\
\mathbb{E}_{\boldsymbol{X},\xi,\boldsymbol{X}'}\Big[\Psi^U_{\mathrm{odd},\ell}\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X},Y_{\max})\Big] &\approx \mathbb{E}_{\boldsymbol{X},\xi_Q,\boldsymbol{X}'}\Big[\Psi^U_{\mathrm{odd},\ell}\mathbf{1}\{T \leq Y_{\max,Q} \leq T+1\}\Big] \\
&= U\,\mathrm{Mat}_r(\mathsf{T}^Q_{2r+1} \lrcorner \boldsymbol{e}_\ell)U^\top \qquad (\ell \in [n]).
\end{aligned} \tag{110}$$

We use the even feature for $2 \leq r \leq s$ and the odd coordinate features at $r = s$.

Algorithm 8.1: EstimateMomentsByPeeling

Input: $\widehat{\mathcal{R}}$; a fresh sample block; a fixed family $(\Psi_\ell)_{\ell\in[m]}$ of symmetric $p \times p$ matrix features evaluated from the supplied samples; accuracy $\varepsilon_{\mathrm{est}} \in (0,1)$; failure probability $\delta \in (0, 1/3)$

Output: Symmetric matrices $\widehat{M}_1, \ldots, \widehat{M}_m$

// Lemma 4.7 and Appendix B.2

$(\widehat{M}_1, \ldots, \widehat{M}_m) \leftarrow \texttt{TrimmedMean}\big(\big(\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X},Y_{\max})\Psi_\ell(\boldsymbol{X},\boldsymbol{X}')\big)_{\ell\in[m]}; 100^{s+2}, \frac{\varepsilon_{\mathrm{est}}}{p\sqrt{m}}, \delta\big)$;

return $(\widehat{M}_1, \ldots, \widehat{M}_m)$;

8.1.3 Error Decomposition

We now decompose the estimation error. Let $\Psi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{p\times p}$ be measurable with $\mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'}[\|\Psi(\boldsymbol{X},\boldsymbol{X}')\|_F] < \infty$, and let $\omega_R(\boldsymbol{X},Y_{\max})$ and $\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X},Y_{\max})$ denote the true and estimated peeling weights from (104) and (106). For a matrix estimate $\widehat{M}$, the error decomposes as follows.

$$\begin{aligned}
&\widehat{M} - \mathbb{E}_{\boldsymbol{X},\xi_Q,\boldsymbol{X}'}[\Psi(\boldsymbol{X},\boldsymbol{X}')\mathbf{1}\{T \leq Y_{\max,Q} \leq T+1\}] \\
&\quad= \underbrace{\widehat{M} - \mathbb{E}_{\boldsymbol{X},Y_{\max},\boldsymbol{X}'}[\Psi(\boldsymbol{X},\boldsymbol{X}')\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X},Y_{\max})]}_{\text{sampling error}} \\
&\qquad+ \underbrace{\mathbb{E}_{\boldsymbol{X},Y_{\max},\boldsymbol{X}'}\big[\Psi(\boldsymbol{X},\boldsymbol{X}')\big(\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X},Y_{\max}) - \omega_R(\boldsymbol{X},Y_{\max})\big)\big]}_{\text{error from recovered regressors}} \\
&\qquad+ \underbrace{\mathbb{E}_{\boldsymbol{X},Y_{\max},\boldsymbol{X}'}[\Psi(\boldsymbol{X},\boldsymbol{X}')\omega_R(\boldsymbol{X},Y_{\max})] - \mathbb{E}_{\boldsymbol{X},\xi_Q,\boldsymbol{X}'}[\Psi(\boldsymbol{X},\boldsymbol{X}')\mathbf{1}\{T \leq Y_{\max,Q} \leq T+1\}]}_{\text{bias due to signed peeling weight}}.
\end{aligned} \tag{111}$$

The three terms in each bound are the sampling error, the error from the recovered regressors, and the bias due to the signed peeling weight, respectively. We now state the bounds for each term.

Sampling error. The first term in the error decomposition (111) is the sampling error in estimating the weighted expectation formed from the recovered list. We bound this error for the even estimate and for every contraction of the odd estimates.

Lemma 8.2 (Sampling Error for Projected Hermite Moments)**.** *Fix integers $2 \le r \le s$, a matrix $U \in \mathbb{R}^{p \times n^r}$ with orthonormal rows, and a recovered list $\widehat{\mathcal{R}}$. Run* EstimateMomentsByPeeling *(Algorithm 8.1) with accuracy $\varepsilon_{\mathrm{est}} \in (0,1)$ and failure probability $\delta \in (0, 1/3)$. Each call uses fresh samples from the Gaussian self-selection model (Definition 1.1), independent of U and $\widehat{\mathcal{R}}$, and an independent standard Gaussian vector $\boldsymbol{X}'$ for each sample.*

(i) Given the even feature Ψ^U_{even}, the algorithm returns $\widehat{M}$ satisfying, with probability at least $1-\delta$,

$$\left\| \widehat{M} - \mathbb{E}_{\boldsymbol{X}, Y_{\max}, \boldsymbol{X}'} \left[\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X}, Y_{\max}) \Psi^U_{\mathrm{even}}(\boldsymbol{X}, \boldsymbol{X}') \right] \right\|_F \le \varepsilon_{\mathrm{est}}.$$

(ii) At $r = s$, given the odd features $(\Psi^U_{\mathrm{odd},\ell})_{\ell \in [n]}$, the algorithm returns $\widehat{M}_1, \dots, \widehat{M}_n$ satisfying, with probability at least $1-\delta$, simultaneously for every $\boldsymbol{b} \in \mathbb{R}^n$,

$$\begin{aligned} &\left\| \sum_{\ell=1}^n b_\ell \left(\widehat{M}_\ell - \mathbb{E}_{\boldsymbol{X}, Y_{\max}, \boldsymbol{X}'} \left[\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X}, Y_{\max}) \Psi^U_{\mathrm{odd},\ell}(\boldsymbol{X}, \boldsymbol{X}') \right] \right) \right\|_F \\ &\qquad \le \varepsilon_{\mathrm{est}} \|\boldsymbol{b}\|_2 . \end{aligned} \tag{112}$$

We prove Lemma 8.2 in Section 8.2. The proof applies a general mean estimation guarantee under an entrywise second-moment bound, which we verify for the weighted even and odd Hermite features.

Recovered-regressor error. The second term in the error decomposition (111) measures the change in the weighted expectation when the true recovered regressors are replaced by their estimates.

For a scalar function ψ, write $\|\psi\|_{L^2(\boldsymbol{X},\boldsymbol{X}')} = (\mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'}[|\psi(\boldsymbol{X},\boldsymbol{X}')|^2])^{1/2}$.

Lemma 8.3 (Peeling-Weight Stability)**.** *Consider the Gaussian self-selection model (Definition 1.1). Fix a list $\widehat{\mathcal{R}}$ of vectors in $\mathbb{R}^n$, a set $R \subsetneq [k]$, and a bijection* match $: \widehat{\mathcal{R}} \to R$. *Let $\boldsymbol{X}' \sim N(\boldsymbol{0}, I_n)$ be independent of the covariate and noises. For every measurable function $\psi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with $\|\psi\|_{L^2(\boldsymbol{X},\boldsymbol{X}')} < \infty$,*

$$\begin{aligned} &\left| \mathbb{E}_{\boldsymbol{X}, Y_{\max}, \boldsymbol{X}'} \left[\psi(\boldsymbol{X}, \boldsymbol{X}') \left(\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X}, Y_{\max}) - \omega_R(\boldsymbol{X}, Y_{\max}) \right) \right] \right| \\ &\quad \le \frac{1}{\sqrt{2\pi}} \|\psi\|_{L^2(\boldsymbol{X},\boldsymbol{X}')} \left(\sum_{\widehat{\boldsymbol{w}} \in \widehat{\mathcal{R}}} \left\| \widehat{\boldsymbol{w}} - \boldsymbol{w}_{\mathrm{match}(\widehat{\boldsymbol{w}})} \right\|_2^2 \right)^{1/2} . \end{aligned} \tag{113}$$

We prove Lemma 8.3 in Section 8.3 using a Gaussian translation bound.

Peeling bias. Even with the true recovered regressors, the signed weight can differ from the slice indicator of the unrecovered maximum. The third term in the error decomposition (111) captures this bias. We identify this bias in the lemma below and bound its magnitude.

Lemma 8.4 (Peeling Bias Bound). *Consider the Gaussian self-selection model (Definition 1.1). Fix $R \subsetneq [k]$, put $Q = [k] \setminus R$, and let $\boldsymbol{X}'$ be a random element independent of the covariate and noises.*

(i) For every integrable scalar- or matrix-valued function $\psi(\boldsymbol{X}, \boldsymbol{X}')$,

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{X},\xi,\boldsymbol{X}'}\left[\psi(\boldsymbol{X},\boldsymbol{X}')\omega_R(\boldsymbol{X}, Y_{\max})\right] - \mathbb{E}_{\boldsymbol{X},\xi,\boldsymbol{X}'}\left[\psi(\boldsymbol{X},\boldsymbol{X}')\mathbf{1}\{T \le Y_{\max,Q} \le T+1\}\right] \\ &\quad = -\mathbb{E}_{\boldsymbol{X},\xi,\boldsymbol{X}'}\left[\psi(\boldsymbol{X},\boldsymbol{X}')\mathbf{1}\{T \le \min\{Y_{\max,R}, Y_{\max,Q}\} \le T+1\}\right]. \end{aligned} \tag{114}$$

(ii) Suppose the boundedness and uncoveredness assumptions (Assumption 1.2) hold, and use the choice of K from Hyperparameters 4.1. Let $s \ge 1$ be an integer and suppose $T > 0$ and $T^2 \ge Ks$. If $\boldsymbol{X}' \sim N(\mathbf{0}, I_n)$ and $\psi(\boldsymbol{X}, \boldsymbol{X}')$ is a real polynomial of degree at most $2s+1$, then for every $1 \le r \le s$,

$$\begin{aligned} &\left|\mathbb{E}_{\boldsymbol{X},\xi,\boldsymbol{X}'}\left[\psi(\boldsymbol{X},\boldsymbol{X}')\mathbf{1}\{T \le \min\{Y_{\max,R}, Y_{\max,Q}\} \le T+1\}\right]\right| \\ &\quad \le k e^{4s\log K - T^2/(2K)} \, \|\psi\|_{L^2(\boldsymbol{X},\boldsymbol{X}')} \operatorname{tr}(A_Q^{(r)}). \end{aligned}$$

The identity in Lemma 8.4(i) follows by conditioning on $\boldsymbol{X}$ in the peeling identity above. We prove Lemma 8.4(ii) in Section 8.4 by applying Hölder's inequality and then bounding the resulting factors using Gaussian hypercontractivity and standard tail bounds.

Finally, we combine all the ingredients from the roadmap to prove Theorem 8.1 in Section 8.5. The rest of this section is dedicated to the proofs of the statements above.

8.2 Sampling Error

To prove Lemma 8.2, we first bound the second moment of every weighted matrix entry, then apply a general mean estimation guarantee. Since the signed peeling weight has absolute value at most one, it suffices to bound the entries of the projected Hermite features.

Recall that $\|\psi\|_{L^2(\boldsymbol{X},\boldsymbol{X}')} = (\mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'}[|\psi(\boldsymbol{X},\boldsymbol{X}')|^2])^{1/2}$.

Proposition 8.5 (Projected-Feature Second Moments). *Let $r \ge 1$, and fix a projection map $U : \mathbb{R}^{n^r} \to \mathbb{R}^p$, where $U \in \mathbb{R}^{p \times n^r}$ has orthonormal rows. Let $\boldsymbol{X}, \boldsymbol{X}' \sim N(\mathbf{0}, I_n)$ be independent. For the projected features Ψ^U_{even} and $(\Psi^U_{\text{odd},\ell})_{\ell\in[n]}$ defined in (107) and (108), a matrix $H \in \mathbb{R}^{p\times p}$ with $\|H\|_F \le 1$, and any $\boldsymbol{b} \in \mathbb{R}^n$, write*

$$\begin{aligned} \psi(\boldsymbol{X},\boldsymbol{X}') &:= \operatorname{tr}\left(H^\top \Psi^U_{\text{even}}(\boldsymbol{X},\boldsymbol{X}')\right), \\ \psi_{\boldsymbol{b}}(\boldsymbol{X},\boldsymbol{X}') &:= \operatorname{tr}\left(H^\top \sum_{\ell=1}^n b_\ell \Psi^U_{\text{odd},\ell}(\boldsymbol{X},\boldsymbol{X}')\right). \end{aligned}$$

These are real polynomials in $(\boldsymbol{X}, \boldsymbol{X}') \in \mathbb{R}^{2n}$ of degrees at most $2r$ and $2r+1$, respectively, and

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'}[\psi(\boldsymbol{X},\boldsymbol{X}')^2] \le 100^{r+2}, \qquad \mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'}[\psi_{\boldsymbol{b}}(\boldsymbol{X},\boldsymbol{X}')^2] \le 100^{r+2}\,\|\boldsymbol{b}\|_2^2.$$

Moreover, their conditional means satisfy

$$\begin{aligned}\left\|\mathbb{E}_{X'}[\psi(X,X') \mid X]\right\|_{L^2(X)} &\le 1,\\ \left\|\mathbb{E}_{X'}[\psi_b(X,X') \mid X]\right\|_{L^2(X)} &\le \|b\|_2 .\end{aligned} \tag{115}$$

The proof of Proposition 8.5 follows from a standard application of Wick's identity (Proposition F.1) and Hermite orthogonality (Proposition 4.5). As such, we defer the details to Appendix F.1.

We use the following mean estimation bound for a family of matrix-valued functions. Its proof applies the coordinatewise trimmed mean to all entries on the same sample block and is deferred to Appendix F.2.

Lemma 8.6 (Mean Estimation with the Peeling Weight)**.** *Let $\varepsilon_{\text{est}} \in (0,1)$ and $\delta \in (0,1/3)$. Fix a recovered list $\widehat{\mathcal{R}}$ and measurable symmetric matrix-valued functions $\Psi_\ell : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^{p\times p}$, $\ell \in [m]$, where $p, m \ge 1$. Let $(X, Y_{\max})$ be a fresh sample from the Gaussian self-selection model (Definition 1.1), and let $X' \sim N(0, I_n)$ be independent of $(X, Y_{\max})$. For the signed peeling weight $\widehat{\omega}_{\widehat{\mathcal{R}}}(X, Y_{\max})$ (106) formed from $\widehat{\mathcal{R}}$, suppose*

$$\mathbb{E}_{X,Y_{\max},X'}\left[\left(\widehat{\omega}_{\widehat{\mathcal{R}}}(X,Y_{\max})(\Psi_\ell(X,X'))_{ac}\right)^2\right] \le 100^{s+2} \qquad (\ell \in [m],\ a,c \in [p]).$$

Given the recovered list, fresh samples, and these functions, EstimateMomentsByPeeling *(Algorithm 8.1) returns symmetric matrices $\widehat{M}_1, \dots, \widehat{M}_m$ such that, with probability at least $1-\delta$, simultaneously for every $b \in \mathbb{R}^m$,*

$$\left\| \sum_{\ell=1}^m b_\ell \left(\widehat{M}_\ell - \mathbb{E}_{X,Y_{\max},X'}\left[\widehat{\omega}_{\widehat{\mathcal{R}}}(X,Y_{\max})\Psi_\ell(X,X')\right]\right)\right\|_F \le \varepsilon_{\text{est}} \|b\|_2 .$$

Its sample complexity is

$$\widetilde{O}\left(\frac{p^2 m}{\varepsilon_{\text{est}}^2} 100^s \log(1/\delta)\right)$$

and its running time is $\text{poly}(n, k, p, m, 100^s, 1/\varepsilon_{\text{est}}, \log(1/\delta))$.

Armed with the necessary second moment bound (Proposition 8.5) to apply Lemma 8.6, we are now ready to analyze the sampling error.

Proof of Lemma 8.2. Take $H = e_a e_c^\top$ in Proposition 8.5. Since the signed weight has absolute value at most one,

$$\begin{aligned}&\mathbb{E}_{X,Y_{\max},X'}\left[\left(\widehat{\omega}_{\widehat{\mathcal{R}}}(X,Y_{\max})(\Psi^U_{\text{even}}(X,X'))_{ac}\right)^2\right]\\ &\quad\le \mathbb{E}_{X,X'}[(\Psi^U_{\text{even}}(X,X'))^2_{ac}] && (|\widehat{\omega}_{\widehat{\mathcal{R}}}(X,Y_{\max})| \le 1)\\ &\quad\le 100^{r+2} \le 100^{s+2}. && (\text{by Proposition 8.5}, r \le s)\end{aligned}$$

Applying Lemma 8.6 with $m = 1$ gives Lemma 8.2(i).

For the odd call, take $r = s$ and $b = e_\ell$ in the same proposition. This gives the required second-moment bound for every entry of every odd coordinate matrix. Applying Lemma 8.6 with $m = n$ gives Lemma 8.2(ii), simultaneously for every contraction. □

8.3 Recovered-Regressor Error

Having controlled the error due to sampling, we turn to the proof of Lemma 8.3, which quantifies the error from replacing the true regressors in the ideal peeling weight with their estimates.

First, we state a useful translation bound of the Gaussian measure without proof.

Proposition 8.7 (Gaussian Translation Bound; [DMR23, Section 2]). *Let $\boldsymbol{G} \sim N(\mathbf{0}, I_p)$, where $p \geq 1$. For every $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}^p$ and measurable $\mathcal{E} \subseteq \mathbb{R}^p$,*

$$|\mathbb{P}(\boldsymbol{G} + \boldsymbol{\mu} \in \mathcal{E}) - \mathbb{P}(\boldsymbol{G} + \boldsymbol{\nu} \in \mathcal{E})| \leq \frac{\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2}{\sqrt{2\pi}}.$$

This is all the tools we need to control the error due to approximate regressor recovery.

Proof of Lemma 8.3. Let $\widehat{\mathcal{R}}$ be a list of recovered regressors corresponding to regressors indexed by R and ψ be a function with $\|\psi\|_{L^2(\boldsymbol{X}, \boldsymbol{X}')} < \infty$. We wish to bound

$$\left|\mathbb{E}_{\boldsymbol{X}, Y_{\max}, \boldsymbol{X}'}\left[\psi(\boldsymbol{X}, \boldsymbol{X}')\left(\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X}, Y_{\max}) - \omega_R(\boldsymbol{X}, Y_{\max})\right)\right]\right|.$$

We first fix the covariate and compare the probabilities of the same slice under the true and estimated Gaussian means. We then integrate this comparison against ψ and apply Cauchy–Schwarz.

If the recovered list is empty, the two peeling weights coincide. Otherwise, fix $\boldsymbol{x} \in \mathbb{R}^n$, and index the mean vectors by $\widehat{w} \in \widehat{\mathcal{R}}$, with

$$\mu_{\widehat{w}} = \left\langle \boldsymbol{w}_{\mathrm{match}(\widehat{w})}, \boldsymbol{x} \right\rangle, \qquad \widehat{\mu}_{\widehat{w}} = \langle \widehat{\boldsymbol{w}}, \boldsymbol{x} \rangle.$$

Let $\boldsymbol{G} \sim N(\mathbf{0}, I_{|R|})$ and take $\mathcal{E} = \{\boldsymbol{y} \in \mathbb{R}^{|R|} : T \leq \max_j y_j \leq T + 1\}$. Conditional on $\boldsymbol{X} = \boldsymbol{x}$, the true and estimated recovered response vectors have the laws of $\boldsymbol{G} + \boldsymbol{\mu}$ and $\boldsymbol{G} + \widehat{\boldsymbol{\mu}}$. Thus

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{G}}(\boldsymbol{G} + \boldsymbol{\mu} \in \mathcal{E}) &= \mathbb{P}_{\boldsymbol{G}}(\max_j (G_j + \mu_j) \leq T + 1) - \mathbb{P}_{\boldsymbol{G}}(\max_j (G_j + \mu_j) \leq T) \\
&= F_R(T + 1 \mid \boldsymbol{x}) - F_R(T \mid \boldsymbol{x}), \\
\mathbb{P}_{\boldsymbol{G}}(\boldsymbol{G} + \widehat{\boldsymbol{\mu}} \in \mathcal{E}) &= \widehat{F}_{\widehat{\mathcal{R}}}(T + 1 \mid \boldsymbol{x}) - \widehat{F}_{\widehat{\mathcal{R}}}(T \mid \boldsymbol{x}).
\end{aligned} \tag{116}$$

Taking the difference of the weights then gives

$$\begin{aligned}
\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{x}, z) - \omega_R(\boldsymbol{x}, z) &= \left(F_R(T + 1 \mid \boldsymbol{x}) - F_R(T \mid \boldsymbol{x})\right) \\
&\quad - \left(\widehat{F}_{\widehat{\mathcal{R}}}(T + 1 \mid \boldsymbol{x}) - \widehat{F}_{\widehat{\mathcal{R}}}(T \mid \boldsymbol{x})\right) \\
&= \mathbb{P}_{\boldsymbol{G}}(\boldsymbol{G} + \boldsymbol{\mu} \in \mathcal{E}) - \mathbb{P}_{\boldsymbol{G}}(\boldsymbol{G} + \widehat{\boldsymbol{\mu}} \in \mathcal{E}). \qquad \text{(by (116))}
\end{aligned}$$

The Gaussian translation bound therefore gives, for every $z \in \mathbb{R}$,

$$\begin{aligned}
\left|\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{x}, z) - \omega_R(\boldsymbol{x}, z)\right| &\leq \frac{\|\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2}{\sqrt{2\pi}} \qquad \text{(by Proposition 8.7)} \\
&= \frac{1}{\sqrt{2\pi}} \left(\sum_{\widehat{w} \in \widehat{\mathcal{R}}} \left\langle \widehat{w} - \boldsymbol{w}_{\mathrm{match}(\widehat{w})}, \boldsymbol{x} \right\rangle^2 \right)^{1/2}.
\end{aligned} \tag{117}$$

We now integrate this pointwise bound against ψ. The difference of the weights does not depend on $Y_{\max}$, so

$$\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{X},Y_{\max},\boldsymbol{X}'}\left[\psi(\boldsymbol{X},\boldsymbol{X}')\left(\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X},Y_{\max})-\omega_R(\boldsymbol{X},Y_{\max})\right)\right]\right| \\
&\quad\le \frac{1}{\sqrt{2\pi}}\mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'}\left[|\psi(\boldsymbol{X},\boldsymbol{X}')|\left(\sum_{\widehat{w}\in\widehat{\mathcal{R}}}\left\langle \widehat{w}-\boldsymbol{w}_{\mathrm{match}(\widehat{w})},\boldsymbol{X}\right\rangle^2\right)^{1/2}\right] && \text{(by (117))} \\
&\quad\le \frac{1}{\sqrt{2\pi}}\|\psi\|_{L^2(\boldsymbol{X},\boldsymbol{X}')}\left(\mathbb{E}_{\boldsymbol{X}}\sum_{\widehat{w}\in\widehat{\mathcal{R}}}\left\langle \widehat{w}-\boldsymbol{w}_{\mathrm{match}(\widehat{w})},\boldsymbol{X}\right\rangle^2\right)^{1/2} && \text{(Cauchy–Schwarz)} \\
&\quad= \frac{1}{\sqrt{2\pi}}\|\psi\|_{L^2(\boldsymbol{X},\boldsymbol{X}')}\left(\sum_{\widehat{w}\in\widehat{\mathcal{R}}}\left\|\widehat{w}-\boldsymbol{w}_{\mathrm{match}(\widehat{w})}\right\|_2^2\right)^{1/2}. && \text{($\mathbb{E}_{\boldsymbol{X}}[\boldsymbol{X}\boldsymbol{X}^\top]=I_n$)}
\end{aligned}$$

This proves the lemma. □

8.4 Peeling Bias

Finally, we prove Lemma 8.4, which gives a bound on the bias resulting from using the peeling weight $\omega_R(\boldsymbol{X},Y_{\max})$ as a proxy for the indicator $\mathbf{1}\{T\le Y_{\max,Q}\le T+1\}$. Our main tool is this pairwise bound stated below whose proof follows from an application of Gaussian hypercontractivity and standard tail bounds to the factors obtained via Hölder's inequality. As such, the proof is deferred to Appendix F.3.

Lemma 8.8 (Pairwise Exceedance Bound)**.** *Consider the Gaussian self-selection model (Definition 1.1) under the boundedness and uncoveredness assumptions (Assumption 1.2). Use the choice of K from Hyperparameters 4.1. Let $s\ge 1$ be an integer and suppose $T>0$ and $T^2\ge Ks$. Let $\boldsymbol{X}\sim N(\mathbf{0},I_n)$ be the covariate and let $\xi_1,\dots,\xi_k\sim N(0,1)$ be independent noises, independent of $\boldsymbol{X}$. Let $\boldsymbol{X}'\sim N(\mathbf{0},I_n)$ be independent of the covariate and noises. Fix a real polynomial $\psi(\boldsymbol{X},\boldsymbol{X}')$ of degree at most $2s+1$. For every pair of distinct indices $i,j\in[k]$ and every $1\le r\le s$,*

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X},\xi_i,\xi_j,\boldsymbol{X}'}\left[|\psi(\boldsymbol{X},\boldsymbol{X}')|\mathbf{1}\{Y_i>T,Y_j>T\}\right] \\
&\quad\le e^{4s\log K-T^2/(2K)}\,\|\psi\|_{L^2(\boldsymbol{X},\boldsymbol{X}')}\,\lambda_j^{(2r)}.
\end{aligned} \tag{118}$$

We are now ready to bound the bias due to the peeling weights.

Proof of Lemma 8.4. For Lemma 8.4(i), fix $R\subsetneq[k]$, put $Q=[k]\setminus R$, and let $\psi(\boldsymbol{X},\boldsymbol{X}')$ be integrable, where $\boldsymbol{X}'$ is independent of the covariate and noises. We have

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X},\xi,\boldsymbol{X}'}\left[\psi\left(\omega_R(\boldsymbol{X},Y_{\max})-\mathbf{1}\{T\le Y_{\max,Q}\le T+1\}\right)\right] \\
&\quad=\mathbb{E}_{\boldsymbol{X},\xi,\boldsymbol{X}'}\left[\psi\begin{pmatrix}\mathbf{1}\{T\le Y_{\max}\le T+1\}-\mathbf{1}\{T\le Y_{\max,R}\le T+1\}\\ -\mathbf{1}\{T\le Y_{\max,Q}\le T+1\}\end{pmatrix}\right] && \text{(conditioning on $\boldsymbol{X},\boldsymbol{X}'$)} \\
&\quad=-\mathbb{E}_{\boldsymbol{X},\xi,\boldsymbol{X}'}\left[\psi\,\mathbf{1}\{T\le\min\{Y_{\max,R},Y_{\max,Q}\}\le T+1\}\right]. && \text{(the max–min identity)}
\end{aligned}$$

This proves the identity, entrywise for a matrix feature.

For Lemma 8.4(ii), we will bound a superset of the indicator event. Indeed, observe that for $\mathbf{1}\{T \le \min\{Y_{\max,R}, Y_{\max,Q}\} \le T+1\} = 1$, it must be that some $i \in R$ and $j \in Q$ satisfy $Y_i > T$ and some $Y_j \ge Y_i > T$, or vice versa. Hence

$$\mathbf{1}\{T \le \min\{Y_{\max,R}, Y_{\max,Q}\} \le T+1\} \le \sum_{i\in R, j\in Q} \mathbf{1}\{Y_i > T, Y_j > T\}.$$

Suppose ψ is a real polynomial of degree at most $2s+1$, $\boldsymbol{X}'$ is standard Gaussian, and $T^2 \ge Ks$. Under the boundedness and uncoveredness assumptions, Lemma 8.8 applies so that

$$\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{X},\xi,\boldsymbol{X}'}\left[\psi(\boldsymbol{X},\boldsymbol{X}')\mathbf{1}\{T \le \min\{Y_{\max,R}, Y_{\max,Q}\} \le T+1\}\right]\right| \\
&\quad\le \sum_{i\in R,\, j\in Q} \mathbb{E}_{\boldsymbol{X},\xi_i,\xi_j,\boldsymbol{X}'}\left[|\psi(\boldsymbol{X},\boldsymbol{X}')|\mathbf{1}\{Y_i > T, Y_j > T\}\right] && \text{(union bound)} \\
&\quad\le e^{4s\log K - T^2/(2K)}\,\|\psi\|_{L^2(\boldsymbol{X},\boldsymbol{X}')}\,|R| \sum_{j\in Q} \lambda_j^{(2r)} && \text{(by Lemma 8.8)} \\
&\quad\le k e^{4s\log K - T^2/(2K)}\,\|\psi\|_{L^2(\boldsymbol{X},\boldsymbol{X}')}\,\mathrm{tr}(A_Q^{(r)}). && (|R| \le k)
\end{aligned}$$

This proves the bound. □

8.5 Proof of Theorem 8.1

We have bounded the sampling error, the error from the recovered regressors, and the peeling bias. We combine these bounds for the even estimate and then for every contraction of the odd estimate.

Proof of Theorem 8.1. Fix a projection map U with orthonormal rows and a list $\widehat{\mathcal{R}}$ of recovered regressors corresponding to regressors indexed by R. Each recovered regressor has estimation error at most $e^{-K^6 s}$. We first prove the even bound at order r, then apply the same argument to each contraction of the odd moment at $r = s$.

Target expectations. We first express the two target matrices as expectations of the projected features restricted to $T \le Y_{\max,Q} \le T+1$. By the definition of the restricted moments and independence of $\boldsymbol{X}'$ from the covariate and latent noises, for every $\boldsymbol{b} \in \mathbb{R}^n$,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X},\xi_Q,\boldsymbol{X}'}[\Psi^U_{\mathrm{even}}(\boldsymbol{X},\boldsymbol{X}')\mathbf{1}\{T \le Y_{\max,Q} \le T+1\}] &= U\,\mathrm{Mat}_r(\mathsf{T}^Q_{2r})U^\top, \\
\mathbb{E}_{\boldsymbol{X},\xi_Q,\boldsymbol{X}'}\left[\left(\sum_{\ell=1}^n b_\ell \Psi^U_{\mathrm{odd},\ell}(\boldsymbol{X},\boldsymbol{X}')\right)\mathbf{1}\{T \le Y_{\max,Q} \le T+1\}\right] &= U\,\mathrm{Mat}_r(\mathsf{T}^Q_{2r+1} \lrcorner \boldsymbol{b})U^\top.
\end{aligned} \tag{119}$$

Even-moment error. We first compare $\widehat{M}$ with the expectation formed from the estimated peeling weight. We then replace the estimated weight by the true weight and the true weight by the slice indicator. By (119), the triangle inequality gives

$$\begin{aligned}
&\left\|\widehat{M} - U\,\mathrm{Mat}_r(\mathsf{T}^Q_{2r})U^\top\right\|_F \\
&\quad\le \left\|\widehat{M} - \mathbb{E}_{\boldsymbol{X},Y_{\max},\boldsymbol{X}'}[\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X},Y_{\max})\Psi^U_{\mathrm{even}}(\boldsymbol{X},\boldsymbol{X}')]\right\|_F && (120) \\
&\qquad + \left\|\mathbb{E}_{\boldsymbol{X},Y_{\max},\boldsymbol{X}'}\left[\left(\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X},Y_{\max}) - \omega_R(\boldsymbol{X},Y_{\max})\right)\Psi^U_{\mathrm{even}}(\boldsymbol{X},\boldsymbol{X}')\right]\right\|_F && (121) \\
&\qquad + \left\|\begin{aligned}&\mathbb{E}_{\boldsymbol{X},Y_{\max},\boldsymbol{X}'}[\omega_R(\boldsymbol{X},Y_{\max})\Psi^U_{\mathrm{even}}(\boldsymbol{X},\boldsymbol{X}')] \\ &\quad - \mathbb{E}_{\boldsymbol{X},\xi_Q,\boldsymbol{X}'}[\Psi^U_{\mathrm{even}}(\boldsymbol{X},\boldsymbol{X}')\mathbf{1}\{T \le Y_{\max,Q} \le T+1\}]\end{aligned}\right\|_F. && (122)
\end{aligned}$$

Sampling error in (120). The first term is the sampling error bounded in Lemma 8.2(i). Thus it is at most ε_{est} with probability at least $1-\delta$.

Recovered-regressor error in (121). The remaining two errors are matrix-valued, while the recovered-regressor and peeling-bias bounds that we proved apply to scalar features. For an error matrix D, we use the following Frobenius duality fact

$$\|D\|_F = \sup_{\|H\|_F \le 1} \left|\operatorname{tr}(H^\top D)\right|.$$

Thus, we fix $H \in \mathbb{R}^{p\times p}$ with $\|H\|_F \le 1$ and bound the pairing uniformly over H. For the even feature, put

$$\bar\psi(\boldsymbol{X}) := \mathbb{E}_{\boldsymbol{X}'}[\operatorname{tr}(H^\top \Psi^U_{\text{even}}(\boldsymbol{X},\boldsymbol{X}')) \mid \boldsymbol{X}].$$

By Proposition 8.5,

$$\deg \bar\psi \le 2r \le 2s, \qquad \|\bar\psi\|_{L^2(\boldsymbol{X})} \le 1.$$

The signed peeling weights are independent of $\boldsymbol{X}'$ conditional on $(\boldsymbol{X}, Y_{\max})$, so the bound from the recovered estimates (Lemma 8.3) gives

$$\begin{aligned}
&\left|\mathbb{E}_{\boldsymbol{X},Y_{\max},\boldsymbol{X}'}\left[\operatorname{tr}(H^\top \Psi^U_{\text{even}}(\boldsymbol{X},\boldsymbol{X}'))\left(\widehat\omega_{\widehat{\mathcal{R}}}(\boldsymbol{X},Y_{\max}) - \omega_R(\boldsymbol{X},Y_{\max})\right)\right]\right| \\
&\quad = \left|\mathbb{E}_{\boldsymbol{X},Y_{\max}}\left[\bar\psi(\boldsymbol{X})\left(\widehat\omega_{\widehat{\mathcal{R}}}(\boldsymbol{X},Y_{\max}) - \omega_R(\boldsymbol{X},Y_{\max})\right)\right]\right| && \text{(averaging over } \boldsymbol{X}'\text{)} \\
&\quad \le \frac{\|\bar\psi\|_{L^2(\boldsymbol{X})}}{\sqrt{2\pi}} \left(\sum_{\widehat w \in \widehat{\mathcal{R}}} \left\|\widehat{\boldsymbol{w}} - \boldsymbol{w}_{\text{match}(\widehat w)}\right\|_2^2\right)^{1/2} && \text{(by Lemma 8.3)} \\
&\quad \le \frac{\sqrt{k}}{\sqrt{2\pi}} e^{-K^6 s}. && \text{(the hypotheses; } \|\bar\psi\|_{L^2(\boldsymbol{X})} \le 1,\ |R| \le k\text{)}
\end{aligned}$$

Taking the supremum over H gives the bound on (121).

Signed-weight bias in (122). It remains to bound the bias from using the true peeling weight in place of the slice indicator. We will again use Frobenius duality. Let H be a unit Frobenius norm matrix and $\bar\psi$ be as above.

$$\begin{aligned}
&\Big|\mathbb{E}_{\boldsymbol{X},Y_{\max}}[\bar\psi(\boldsymbol{X})\omega_R(\boldsymbol{X},Y_{\max})] \\
&\qquad - \mathbb{E}_{\boldsymbol{X},\xi_Q}[\bar\psi(\boldsymbol{X})\mathbf{1}\{T \le Y_{\max,Q} \le T+1\}]\Big| \\
&\quad \le k e^{4s\log K - T^2/(2K)}\, \|\bar\psi\|_{L^2(\boldsymbol{X})}\, \operatorname{tr}(A_Q^{(r)}) && \text{(by Lemmas 8.4(i) and 8.4(ii))} \\
&\quad \le k e^{4s\log K - T^2/(2K)}\, \operatorname{tr}(A_Q^{(r)}). && (\|\bar\psi\|_{L^2(\boldsymbol{X})} \le 1)
\end{aligned}$$

Taking the same supremum bounds (122). Adding the three bounds gives, with probability at least $1-\delta$,

$$\left\|\widehat M - U\,\mathrm{Mat}_r(\mathsf{T}^Q_{2r})U^\top\right\|_F \le \varepsilon_{\text{est}} + \frac{\sqrt{k}}{\sqrt{2\pi}} e^{-K^6 s} + k e^{4s\log K - T^2/(2K)}\operatorname{tr}(A_Q^{(r)}).$$

This proves Theorem 8.1(i).

Odd-moment error. At $r = s$, we apply the even-case decomposition to each contraction of the odd moment. For $\boldsymbol{b} \in \mathbb{R}^n$, the estimator is $\sum_{\ell=1}^n b_\ell \widehat{M}_\ell$, and the second identity in (119) gives its estimation target.

By Lemma 8.2(ii), the sampling error is at most $\varepsilon_{\text{est}} \|\boldsymbol{b}\|_2$ simultaneously for every $\boldsymbol{b} \in \mathbb{R}^n$, with probability at least $1 - \delta$.

For the two deterministic errors, fix $\boldsymbol{b} \in \mathbb{R}^n$. As in the even case, we pair each error matrix with an arbitrary $H \in \mathbb{R}^{p \times p}$ satisfying $\|H\|_F \le 1$. The scalar feature for this contraction is

$$\bar{\psi}_{\boldsymbol{b}}(\boldsymbol{X}) := \mathbb{E}_{\boldsymbol{X}'}\left[\operatorname{tr}\left(H^\top \sum_{\ell=1}^n b_\ell \Psi^U_{\text{odd},\ell}(\boldsymbol{X}, \boldsymbol{X}')\right) \mid \boldsymbol{X}\right].$$

By Proposition 8.5,

$$\deg \bar{\psi}_{\boldsymbol{b}} \le 2s + 1, \qquad \|\bar{\psi}_{\boldsymbol{b}}\|_{L^2(\boldsymbol{X})} \le \|\boldsymbol{b}\|_2 .$$

For the recovered-regressor error, Lemma 8.3 yields the following bound

$$\frac{\|\bar{\psi}_{\boldsymbol{b}}\|_{L^2(\boldsymbol{X})}}{\sqrt{2\pi}} \left(\sum_{\widehat{w} \in \widehat{\mathcal{R}}} \left\|\widehat{w} - \boldsymbol{w}_{\text{match}(\widehat{w})}\right\|_2^2\right)^{1/2} \le \frac{\sqrt{k}}{\sqrt{2\pi}} e^{-K^6 s} \|\boldsymbol{b}\|_2 .$$

For the peeling bias, $\deg \bar{\psi}_{\boldsymbol{b}} \le 2s + 1$, so Lemmas 8.4(i) and 8.4(ii) yield the following bound

$$k e^{4s \log K - T^2/(2K)} \operatorname{tr}(A_Q^{(s)}) \|\bar{\psi}_{\boldsymbol{b}}\|_{L^2(\boldsymbol{X})} \le k e^{4s \log K - T^2/(2K)} \operatorname{tr}(A_Q^{(s)}) \|\boldsymbol{b}\|_2 .$$

Both bounds hold for every H with $\|H\|_F \le 1$. Taking the supremum in each and summing all error bounds gives

$$\left\|\sum_{\ell=1}^n b_\ell \widehat{M}_\ell - U \operatorname{Mat}_s(\mathsf{T}^Q_{2s+1} \lrcorner \boldsymbol{b}) U^\top\right\|_F \le \left(\varepsilon_{\text{est}} + \frac{\sqrt{k}}{\sqrt{2\pi}} e^{-K^6 s} + k e^{4s \log K - T^2/(2K)} \operatorname{tr}(A_Q^{(s)})\right) \|\boldsymbol{b}\|_2 .$$

Thus the estimator satisfies Theorem 8.1(ii) for every $\boldsymbol{b} \in \mathbb{R}^n$, with probability at least $1 - \delta$.

Sample complexity and running time. The even call estimates $m = 1$ matrix, and the odd call estimates $m = n$ matrices on the same sample block. By Lemma 8.6, each call uses

$$\widetilde{O}\left(\frac{p^2 m}{\varepsilon_{\text{est}}^2} 100^s \log(1/\delta)\right)$$

projected samples. Evaluating the features and weights and computing the trimmed means takes the sample count times $\text{poly}(n, k, p, m)$ operations. Since $s = K^4 \log n$, this gives the stated running time. □

This gives the moment estimation oracle used at each recursive step of Section 6.2.

9 Proof of Theorem 1.3

We have shown how to obtain rough estimates for a subset of normalized regressors through implicit tensor decomposition (Theorem 6.1) and refine these unit vectors into accurate estimates of the corresponding regressors using a variant of OLS (Theorem 7.1). We now combine these guarantees to prove Theorem 1.3. Our full algorithm simply repeats the two aforementioned subroutines until all regressors are recovered. Its pseudocode can be found in Algorithm 9.1.

Algorithm 9.1: RECOVERALLREGRESSORS

Input: Known $B, \Delta, k, \varepsilon, \delta$, and sample access to $(\boldsymbol{X}, Y_{\max})$
Output: k vectors in $\mathbb{R}^n$

Use fresh observations in every call;
Choose $K, s, T, \varepsilon_{\text{int}}, \Gamma$ as in Hyperparameters 4.1;
$\widehat{\mathcal{R}} \leftarrow ()$;
while $|\widehat{\mathcal{R}}| < k$ **do**
    // Algorithm 6.4 and Theorem 6.1
    $\widehat{\mathcal{D}} \leftarrow \texttt{ImplicitTensorDecomposition}(\widehat{\mathcal{R}}, \frac{\delta}{2k})$;
    foreach $\widehat{\boldsymbol{u}} \in \widehat{\mathcal{D}}$ **do**
        // Algorithm 7.1 and Theorem 7.1
        $\widehat{\boldsymbol{w}} \leftarrow \texttt{RestrictedLeastSquares}(\widehat{\boldsymbol{u}}, \varepsilon_{\text{int}}, \frac{\delta}{2k})$;
        Append $\widehat{\boldsymbol{w}}$ to $\widehat{\mathcal{R}}$;
return $\widehat{\mathcal{R}}$;

We now analyze the guarantees of Algorithm 9.1, which yields a proof of our main theorem.

Proof of Theorem 1.3. We first show that each round appends estimates of previously unrecovered regressors with the accuracy needed for the next decomposition call. We then count the calls to bound the failure probability, sample complexity, and running time.

Accuracy after each round. We first prove correctness when the decomposition and refinement calls satisfy their guarantees. At the beginning of each round, assume we recovered a subset of regressors indexed by $R \subseteq [k]$, with estimates $\widehat{\mathcal{R}}$ up to error ε_{int}:

$$\left\| \widehat{\boldsymbol{w}} - \boldsymbol{w}_{\text{match}(\widehat{\boldsymbol{w}})} \right\|_2 \le \varepsilon_{\text{int}} \qquad (\widehat{\boldsymbol{w}} \in \widehat{\mathcal{R}}).$$

Here $\text{match}(\widehat{\boldsymbol{w}})$ denotes the index of the true regressor corresponding to $\widehat{\boldsymbol{w}}$. This vacuously holds for the initial empty list.

Fix a current list satisfying this property, with $R \neq [k]$. By the choice of ε_{int} in Hyperparameters 4.1, $\varepsilon_{\text{int}} \le e^{-K^6 s}$. Thus $\widehat{\mathcal{R}}$ satisfies the input accuracy requirement of IMPLICITTENSORDECOMPOSITION (Theorem 6.1). By Theorem 6.1, IMPLICITTENSORDECOMPOSITION returns a list $\widehat{\mathcal{D}}$ of unit vectors matched to some normalized regressors indexed by a non-empty subset $J \subseteq [k] \setminus R$, with errors at most $\Delta^2/(16B^2)$. Execute RESTRICTEDLEASTSQUARES on each $\widehat{\boldsymbol{u}} \in \widehat{\mathcal{D}}$ with requested accuracy ε_{int}. Theorem 7.1 guarantees that the subroutine returns ε_{int}-accurate estimates $\widehat{\boldsymbol{w}}_i$ for regressors indexed by J.

Completion and failure probability. Each round adds at least one previously unmatched regressor, so there are at most k calls to ImplicitTensorDecomposition. Since each regressor is refined once, there are also at most k calls to RestrictedLeastSquares. At termination, the matching gives a permutation π of $[k]$ such that

$$\max_{i \in [k]} \left\| \widehat{\boldsymbol{w}}_i - \boldsymbol{w}_{\pi(i)} \right\|_2 \leq \varepsilon_{\text{int}} \leq \varepsilon.$$

By a union bound over these at most $2k$ calls, each using fresh observations and failure probability $\delta/(2k)$, the algorithm succeeds with probability at least $1 - \delta$.

Sample complexity and running time. There are at most k calls to each routine. Each call uses failure probability $\delta/(2k)$, and each refinement call uses accuracy ε_{int}. By the sample complexity and running time guarantees of Theorems 6.1 and 7.1, the total sample complexity is at most

$$\widetilde{O}\left(k \left[(nk)^{4K^{10}+12K^4} + \left(\frac{nk}{\varepsilon_{\text{int}}} \right)^{5K^2} \right] \log\left(\frac{2k}{\delta} \right) \right) = \widetilde{O}\left(\left(\frac{nk}{\varepsilon} \right)^{K^{13}} \log(1/\delta) \right),$$

and the total running time is at most

$$\widetilde{O}\left(k \left[(nk)^{\text{poly}(K)} + \left(\frac{nk}{\varepsilon_{\text{int}}} \right)^{\text{poly}(K)} \right] \log\left(\frac{2k}{\delta} \right) \right) = \left(\frac{nk}{\varepsilon} \right)^{\text{poly}(K)} \log(1/\delta).$$

The equality is due to the choice of Hyperparameters 4.1 and some mundane simplifications deferred to the appendix (Proposition G.1). □

References

[AGHK+14] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. "Tensor Decompositions for Learning Latent Variable Models". In: *Journal of Machine Learning Research* 15.80 (2014), pp. 2773–2832 (cit. on pp. 11, 14).

[AH02] Susan Athey and Philip A. Haile. "Identification of Standard Auction Models". In: *Econometrica* 70.6 (2002), pp. 2107–2140 (cit. on p. 1).

[AHK12] Animashree Anandkumar, Daniel Hsu, and Sham M. Kakade. "A Method of Moments for Mixture Models and Hidden Markov Models". In: *Proceedings of the 25th Annual Conference on Learning Theory*. Vol. 23. Proceedings of Machine Learning Research. PMLR, 2012, pp. 33.1–33.34 (cit. on pp. 3, 8, 14, 38, 96).

[AK05] Sanjeev Arora and Ravi Kannan. "Learning Mixtures of Separated Nonspherical Gaussians". In: *The Annals of Applied Probability* 15.1A (2005), pp. 69–92 (cit. on p. 14).

[AM05] Dimitris Achlioptas and Frank McSherry. "On Spectral Learning of Mixtures of Distributions". In: *Learning Theory: 18th Annual Conference on Learning Theory*. Vol. 3559. Lecture Notes in Computer Science. Springer, 2005, pp. 458–469 (cit. on p. 13).

[BCMV14] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. "Smoothed Analysis of Tensor Decompositions". In: *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*. Association for Computing Machinery, 2014, pp. 594–603 (cit. on p. 96).

[BDJK+22] Ainesh Bakshi, Ilias Diakonikolas, He Jia, Daniel M. Kane, Pravesh K. Kothari, and Santosh S. Vempala. "Robustly Learning Mixtures of k Arbitrary Gaussians". In: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*. Association for Computing Machinery, 2022, pp. 1234–1247 (cit. on p. 14).

[BGS15] Gábor Balázs, András György, and Csaba Szepesvári. "Near-Optimal Max-Affine Estimators for Convex Regression". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Vol. 38. Proceedings of Machine Learning Research. PMLR, 2015, pp. 56–64 (cit. on p. 13).

[Bha97] Rajendra Bhatia. *Matrix Analysis*. Vol. 169. Graduate Texts in Mathematics. Springer, 1997 (cit. on pp. 17, 76, 83).

[BS10] Mikhail Belkin and Kaushik Sinha. "Toward Learning Gaussian Mixtures with Arbitrary Separation". In: *Proceedings of the 23rd Annual Conference on Learning Theory*. 2010, pp. 407–419 (cit. on p. 14).

[BWY17] Sivaraman Balakrishnan, Martin J. Wainwright, and Bin Yu. "Statistical Guarantees for the EM Algorithm: From Population to Sample-Based Analysis". In: *The Annals of Statistics* 45.1 (2017), pp. 77–120 (cit. on p. 14).

[CDGK+23] Sitan Chen, Zehao Dou, Surbhi Goel, Adam R. Klivans, and Raghu Meka. "Learning Narrow One-Hidden-Layer ReLU Networks". In: *Proceedings of Thirty Sixth Conference on Learning Theory*. Vol. 195. Proceedings of Machine Learning Research. PMLR, 2023, pp. 5580–5614 (cit. on p. 14).

[CDIZ23] Yeshwanth Cherapanamjeri, Constantinos Daskalakis, Andrew Ilyas, and Manolis Zampetakis. "What Makes a Good Fisherman? Linear Regression under Self-Selection Bias". In: *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*. Association for Computing Machinery, 2023, pp. 1699–1712 (cit. on pp. i, 1, 2, 4, 11, 13, 96).

[CL13] Arun Tejasvi Chaganty and Percy Liang. "Spectral Experts for Estimating Mixtures of Linear Regressions". In: *Proceedings of the 30th International Conference on Machine Learning*. Vol. 28. Proceedings of Machine Learning Research 3. PMLR, 2013, pp. 1040–1048 (cit. on p. 14).

[CLS20] Sitan Chen, Jerry Li, and Zhao Song. "Learning Mixtures of Linear Regressions in Subexponential Time via Fourier Moments". In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. Association for Computing Machinery, 2020, pp. 587–600 (cit. on p. 14).

[CM20] Sitan Chen and Raghu Meka. "Learning Polynomials in Few Relevant Dimensions". In: *Proceedings of Thirty Third Conference on Learning Theory*. Vol. 125. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1161–1227 (cit. on p. 14).

[CN24] Sitan Chen and Shyam Narayanan. "A Faster and Simpler Algorithm for Learning Shallow Networks". In: *Proceedings of Thirty Seventh Conference on Learning Theory*. Vol. 247. Proceedings of Machine Learning Research. PMLR, 2024, pp. 981–994 (cit. on p. 14).

[Das99] Sanjoy Dasgupta. "Learning Mixtures of Gaussians". In: *Proceedings of the 40th Annual Symposium on Foundations of Computer Science*. IEEE Computer Society, 1999, pp. 634–644 (cit. on p. 13).

[De 89] Richard D. De Veaux. "Mixtures of Linear Regressions". In: *Computational Statistics & Data Analysis* 8.3 (1989), pp. 227–245 (cit. on p. 14).

[DGTZ18] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. "Efficient Statistics, in High Dimensions, from Truncated Samples". In: *2018 IEEE 59th Annual Symposium on Foundations of Computer Science*. IEEE, 2018, pp. 639–649 (cit. on p. 14).

[DGTZ19] Constantinos Daskalakis, Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. "Computationally and Statistically Efficient Truncated Regression". In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Vol. 99. Proceedings of Machine Learning Research. PMLR, 2019, pp. 955–960 (cit. on p. 14).

[DIKR25] Ilias Diakonikolas, Giannis Iakovidis, Daniel M. Kane, and Lisheng Ren. "Algorithms and SQ Lower Bounds for Robustly Learning Real-Valued Multi-Index Models". In: *Advances in Neural Information Processing Systems*. Vol. 38. 2025 (cit. on p. 14).

[DIKZ25] Ilias Diakonikolas, Giannis Iakovidis, Daniel M. Kane, and Nikos Zarifis. "Robust Learning of Multi-index Models via Iterative Subspace Approximation". In: *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2025, pp. 2231–2239 (cit. on p. 14).

[DK20] Ilias Diakonikolas and Daniel M. Kane. "Small Covers for Near-Zero Sets of Polynomials and Learning Latent Variable Models". In: *2020 IEEE 61st Annual Symposium on Foundations of Computer Science*. IEEE, 2020, pp. 184–195 (cit. on p. 14).

[DK24] Ilias Diakonikolas and Daniel M. Kane. "Efficiently Learning One-Hidden-Layer ReLU Networks via Schur Polynomials". In: *Proceedings of the Thirty-Seventh Conference on Learning Theory*. Vol. 247. Proceedings of Machine Learning Research. PMLR, 2024, pp. 1364–1378 (cit. on pp. 14, 16).

[DK25] Ilias Diakonikolas and Daniel M. Kane. "Implicit High-Order Moment Tensor Estimation and Learning Latent Variable Models". In: *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2025, pp. 1228–1247 (cit. on pp. 3, 7, 14, 32, 96).

[DKKZ20] Ilias Diakonikolas, Daniel M. Kane, Vasilis Kontonis, and Nikos Zarifis. "Algorithms and SQ Lower Bounds for PAC Learning One-Hidden-Layer ReLU Networks". In: *Proceedings of the Thirty-Third Conference on Learning Theory*. Vol. 125. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1514–1539 (cit. on pp. 14, 16).

[DKM09] Sanjoy Dasgupta, Adam Tauman Kalai, and Claire Monteleoni. "Analysis of Perceptron-Based Active Learning". In: *Journal of Machine Learning Research* 10 (2009), pp. 281–299 (cit. on p. 87).

[DKPZ24] Ilias Diakonikolas, Daniel M. Kane, Thanasis Pittas, and Nikos Zarifis. "Statistical Query Lower Bounds for Learning Truncated Gaussians". In: *Proceedings of the Thirty-Seventh Conference on Learning Theory*. Vol. 247. Proceedings of Machine Learning Research. PMLR, 2024, pp. 1336–1363 (cit. on p. 15).

[DLB25] Alex Damian, Jason D. Lee, and Joan Bruna. "The Generative Leap: Tight Sample Complexity for Efficiently Learning Gaussian Multi-Index Models". In: *Advances in Neural Information Processing Systems*. Vol. 38. 2025 (cit. on p. 14).

[DLMF] National Institute of Standards and Technology. *NIST Digital Library of Mathematical Functions* (cit. on pp. 16, 22).

[DLNS24] Anindya De, Huan Li, Shivam Nadimpalli, and Rocco A. Servedio. "Detecting Low-Degree Truncation". In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. Association for Computing Machinery, 2024, pp. 1027–1038 (cit. on p. 15).

[DLS22] Alexandru Damian, Jason D. Lee, and Mahdi Soltanolkotabi. "Neural Networks can Learn Representations with Gradient Descent". In: *Proceedings of the Thirty-Fifth Conference on Learning Theory*. Vol. 178. Proceedings of Machine Learning Research. PMLR, 2022, pp. 5413–5452 (cit. on p. 14).

[DMR23] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. *The total variation distance between high-dimensional Gaussians with the same mean*. arXiv:1810.08693v7. 2023 (cit. on p. 61).

[DPLB24] Alex Damian, Loucas Pillaud-Vivien, Jason D. Lee, and Joan Bruna. "Computational-Statistical Gaps in Gaussian Single-Index Models (Extended Abstract)". In: *Proceedings of Thirty Seventh Conference on Learning Theory*. Vol. 247. Proceedings of Machine Learning Research. PMLR, 2024, pp. 1262–1262 (cit. on p. 14).

[DRZ20] Constantinos Daskalakis, Dhruv Rohatgi, and Emmanouil Zampetakis. "Truncated Linear Regression in High Dimensions". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020 (cit. on p. 14).

[FJ72] Ray C. Fair and Dwight M. Jaffee. "Methods of Estimation for Markets in Disequilibrium". In: *Econometrica* 40.3 (1972), pp. 497–514 (cit. on p. 1).

[FKKT21] Dimitris Fotakis, Alkis Kalavasis, Vasilis Kontonis, and Christos Tzamos. "Efficient Algorithms for Learning from Coarse Labels". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Vol. 134. Proceedings of Machine Learning Research. PMLR, 2021, pp. 2060–2079 (cit. on p. 14).

[GHK15] Rong Ge, Qingqing Huang, and Sham M. Kakade. "Learning Mixtures of Gaussians in High Dimensions". In: *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*. Association for Computing Machinery, 2015, pp. 761–770 (cit. on p. 14).

[GM24] Jason Gaitonde and Elchanan Mossel. "Sample-Efficient Linear Regression with Self-Selection Bias". In: *arXiv preprint arXiv:2402.14229* (2024) (cit. on pp. i, 2, 4, 10, 13, 96).

[GN80] Sander Greenland and Raymond Neutra. "Control of Confounding in the Assessment of Medical Technology". In: *International Journal of Epidemiology* 9.4 (1980), pp. 361–367 (cit. on p. 1).

[GPGR22] Avishek Ghosh, Ashwin Pananjady, Adityanand Guntuboyina, and Kannan Ramchandran. "Max-Affine Regression: Parameter Estimation for Gaussian Designs". In: *IEEE Transactions on Information Theory* 68.3 (2022), pp. 1851–1885 (cit. on p. 13).

[Gro81] Sanford J. Grossman. "The Informational Role of Warranties and Private Disclosure about Product Quality". In: *The Journal of Law and Economics* 24.3 (1981), pp. 461–483 (cit. on p. 1).

[GVX14] Navin Goyal, Santosh Vempala, and Ying Xiao. "Fourier PCA and Robust Tensor Decomposition". In: *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*. Association for Computing Machinery, 2014, pp. 584–593 (cit. on p. 96).

[Han22] Bruce E. Hansen. *Probability and Statistics for Economists*. Princeton University Press, 2022 (cit. on p. 75).

[Hig86] Nicholas J. Higham. "Computing the Polar Decomposition—with Applications". In: *SIAM Journal on Scientific and Statistical Computing* 7.4 (1986), pp. 1160–1174 (cit. on p. 83).

[HK13] Daniel Hsu and Sham M. Kakade. "Learning Mixtures of Spherical Gaussians: Moment Methods and Spectral Decompositions". In: *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*. Association for Computing Machinery, 2013, pp. 11–20 (cit. on pp. 3, 8, 14, 96).

[HKZ12] Daniel Hsu, Sham M. Kakade, and Tong Zhang. "A Spectral Algorithm for Learning Hidden Markov Models". In: *Journal of Computer and System Sciences* 78.5 (2012), pp. 1460–1480 (cit. on p. 14).

[HL18] Samuel B. Hopkins and Jerry Li. "Mixture Models, Robustness, and Sum of Squares Proofs". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. Association for Computing Machinery, 2018, pp. 1021–1034 (cit. on p. 14).

[Iss18] L. Isserlis. "On a Formula for the Product-Moment Coefficient of Any Order of a Normal Frequency Distribution in Any Number of Variables". In: *Biometrika* 12.1-2 (1918), pp. 134–139 (cit. on p. 90).

[IZD20] Andrew Ilyas, Emmanouil Zampetakis, and Constantinos Daskalakis. "A Theoretical and Practical Framework for Regression and Classification from Truncated Samples". In: *Proceedings of the Twenty-Third International Conference on Artificial Intelligence and Statistics*. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 4463–4473 (cit. on p. 14).

[Jae00] Herbert Jaeger. "Observable Operator Models for Discrete Stochastic Time Series". In: *Neural Computation* 12.6 (2000), pp. 1371–1398 (cit. on p. 14).

[JSA15] Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. "Score Function Features for Discriminative Learning". In: *3rd International Conference on Learning Representations (ICLR 2015), Workshop Track*. 2015 (cit. on p. 14).

[KBL24] Seonho Kim, Sohail Bahmani, and Kiryung Lee. "Max-Linear Regression by Convex Programming". In: *IEEE Transactions on Information Theory* 70.3 (2024), pp. 1897–1912 (cit. on p. 13).

[KC20] Jeongyeol Kwon and Constantine Caramanis. "EM Converges for a Mixture of Many Linear Regressions". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Vol. 108. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1727–1736 (cit. on p. 14).

[KCL15] Volodymyr Kuleshov, Arun Chaganty, and Percy Liang. "Tensor Factorization via Matrix Factorization". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Vol. 38. Proceedings of Machine Learning Research. PMLR, 2015, pp. 507–516 (cit. on p. 96).

[KKL26] Haitham Kanj, Seonho Kim, and Kiryung Lee. "Sparse Max-Affine Regression". In: *SIAM Journal on Mathematics of Data Science* 8.4 (2026), pp. 1155–1178 (cit. on p. 13).

[KL24] Seonho Kim and Kiryung Lee. "Max-Affine Regression via First-Order Methods". In: *SIAM Journal on Mathematics of Data Science* 6.2 (2024), pp. 534–552 (cit. on p. 13).

[KLMZ26] Vikram Kher, Jane H. Lee, Anay Mehrotra, and Manolis Zampetakis. "Learning-Enabled Estimation: Tight Characterizations under Sample Selection Biases". In: *Proceedings of the 27th ACM Conference on Economics and Computation*. Association for Computing Machinery, 2026 (cit. on p. 15).

[KMKC26] Alexandros Kouridakis, Anay Mehrotra, Alkis Kalavasis, and Constantine Caramanis. "Linear Regression with Unknown Truncation Beyond Gaussian Features". In: *Proceedings of the 43rd International Conference on Machine Learning*. Vol. 306. Proceedings of Machine Learning Research. PMLR, 2026, pp. 60603–60660 (cit. on p. 14).

[KMV10] Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. "Efficiently Learning Mixtures of Two Gaussians". In: *Proceedings of the 42nd ACM Symposium on Theory of Computing*. Association for Computing Machinery, 2010, pp. 553–562 (cit. on p. 14).

[KMZ26] Alkis Kalavasis, Anay Mehrotra, and Felix Zhou. "Can SGD Select Good Fishermen? Local Convergence under Self-Selection Biases (Extended Abstract)". In: *Proceedings of the Thirty-Ninth Conference on Learning Theory*. Vol. 336. Proceedings of Machine Learning Research. PMLR, 2026, pp. 3904–3905 (cit. on pp. i, 2, 4, 11, 13, 14, 96).

[KMZZ+26] Alkis Kalavasis, Anay Mehrotra, Manolis Zampetakis, Felix Zhou, and Ziyu Zhu. "Mean Estimation from Coarse Data: Characterizations and Efficient Algorithms". In: *The Fourteenth International Conference on Learning Representations*. 2026 (cit. on p. 14).

[KSS18] Pravesh K. Kothari, Jacob Steinhardt, and David Steurer. "Robust Moment Estimation and Improved Clustering via Sum of Squares". In: *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. Association for Computing Machinery, 2018, pp. 1035–1046 (cit. on p. 14).

[KTZ19] Vasilis Kontonis, Christos Tzamos, and Manolis Zampetakis. "Efficient Truncated Statistics with Unknown Truncation". In: *2019 IEEE 60th Annual Symposium on Foundations of Computer Science*. IEEE, 2019, pp. 1578–1595 (cit. on p. 14).

[Lan57] H. O. Lancaster. "Some properties of the bivariate normal distribution considered in the form of a contingency table". In: *Biometrika* 44.1-2 (1957), pp. 289–292 (cit. on pp. 5, 96).

[Li91] Ker-Chau Li. "Sliced Inverse Regression for Dimension Reduction". In: *Journal of the American Statistical Association* 86.414 (1991), pp. 316–327 (cit. on p. 14).

[Li92] Ker-Chau Li. "On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma". In: *Journal of the American Statistical Association* 87.420 (1992), pp. 1025–1039 (cit. on p. 14).

[LL18] Yuanzhi Li and Yingyu Liang. "Learning Mixtures of Linear Regressions with Nearly Optimal Complexity". In: *Proceedings of the 31st Conference on Learning Theory*. Vol. 75. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1125–1144 (cit. on p. 14).

[LL24] Jerry Li and Allen Liu. "Clustering Mixtures with Almost Optimal Separation in Polynomial Time". In: *SIAM Journal on Computing* (2024). Special section on STOC 2022, STOC22-46–STOC22-102 (cit. on pp. 3, 7, 11, 14, 32, 96).

[LM22] Allen Liu and Ankur Moitra. "Learning GMMs with Nearly Optimal Robustness Guarantees". In: *Proceedings of the 35th Conference on Learning Theory*. Vol. 178. Proceedings of Machine Learning Research. PMLR, 2022, pp. 2815–2895 (cit. on p. 14).

[LMZ24] Jane H. Lee, Anay Mehrotra, and Manolis Zampetakis. "Efficient Statistics With Unknown Truncation, Polynomial Time Algorithms, Beyond Gaussians". In: *2024 IEEE 65th Annual Symposium on Foundations of Computer Science*. IEEE, 2024, pp. 988–1006 (cit. on p. 14).

[LMZ26] Jane H. Lee, Anay Mehrotra, and Manolis Zampetakis. "Smoothed Analysis of Learning from Positive Samples". In: *Proceedings of the 58th Annual ACM Symposium on Theory of Computing*. Association for Computing Machinery, 2026, pp. 42–53 (cit. on p. 14).

[LRA93] S. E. Leurgans, R. T. Ross, and R. B. Abel. "A Decomposition for Three-Way Arrays". In: *SIAM Journal on Matrix Analysis and Applications* 14.4 (1993), pp. 1064–1083 (cit. on pp. 8, 96).

[LSSW26] Haitong Liu, Deepak Narayanan Sridharan, David Steurer, and Manuel Wiedmer. "Fast Algorithms for Learning a Gaussian under Halfspace Truncation with Optimal Sample Complexity". In: *Proceedings of the Thirty-Ninth Conference on Learning Theory*. Vol. 336. Proceedings of Machine Learning Research. PMLR, 2026, pp. 4724–4818 (cit. on p. 14).

[MB09] Alessandro Magnani and Stephen P. Boyd. "Convex Piecewise-Linear Fitting". In: *Optimization and Engineering* 10.1 (2009), pp. 1–17 (cit. on p. 13).

[Mil81] Paul R. Milgrom. "Good News and Bad News: Representation Theorems and Applications". In: *The Bell Journal of Economics* 12.2 (1981), pp. 380–391 (cit. on p. 1).

[MV10] Ankur Moitra and Gregory Valiant. "Settling the Polynomial Learnability of Mixtures of Gaussians". In: *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE, 2010, pp. 93–102 (cit. on p. 14).

[NR00] Andrew Y. Ng and Stuart J. Russell. "Algorithms for Inverse Reinforcement Learning". In: *Proceedings of the Seventeenth International Conference on Machine Learning*. Morgan Kaufmann, 2000, pp. 663–670 (cit. on p. 1).

[ODo14] Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014 (cit. on pp. 16, 17, 79).

[OOR25] Roberto I. Oliveira, Paulo Orenstein, and Zoraida F. Rico. *Finite-sample properties of the trimmed mean*. arXiv:2501.03694v1. 2025 (cit. on pp. 17, 77).

[Roy51] A. D. Roy. "Some Thoughts on the Distribution of Earnings". In: *Oxford Economic Papers* 3.2 (1951), pp. 135–146 (cit. on p. 1).

[RR83] Paul R. Rosenbaum and Donald B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects". In: *Biometrika* 70.1 (1983), pp. 41–55 (cit. on p. 1).

[RS08] Alvin C. Rencher and G. Bruce Schaalje. *Linear Models in Statistics*. 2nd ed. John Wiley & Sons, 2008 (cit. on p. 75).

[RV17] Oded Regev and Aravindan Vijayaraghavan. "On Learning Mixtures of Well-Separated Gaussians". In: *2017 IEEE 58th Annual Symposium on Foundations of Computer Science*. IEEE, 2017, pp. 85–96 (cit. on p. 14).

[RW06] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006 (cit. on p. 16).

[Sev26] Thomas A. Severini. "A Note on the Application of Hermite Polynomials to Hidden Truncation Models". In: *Journal of Statistical Theory and Practice* 20, 17 (2026) (cit. on pp. 5, 96).

[SJA16] Hanie Sedghi, Majid Janzamin, and Anima Anandkumar. "Provable Tensor Methods for Learning Mixtures of Generalized Linear Models". In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. Vol. 51. Proceedings of Machine Learning Research. PMLR, 2016, pp. 1223–1231 (cit. on p. 14).

[Tob58] James Tobin. "Estimation of Relationships for Limited Dependent Variables". In: *Econometrica* 26.1 (1958), pp. 24–36 (cit. on p. 14).

[Vem10] Santosh S. Vempala. "Learning Convex Concepts from Gaussian Distributions with PCA". In: *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE Computer Society, 2010, pp. 124–130 (cit. on p. 14).

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. 1st ed. Cambridge University Press, 2018 (cit. on p. 75).

[VW04] Santosh S. Vempala and Grant Wang. "A Spectral Algorithm for Learning Mixture Models". In: *Journal of Computer and System Sciences* 68.4 (2004), pp. 841–860 (cit. on p. 13).

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Vol. 48. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019 (cit. on p. 16).

[WN14] Christopher S. Withers and Saralees Nadarajah. "Some conditional and unconditional expectation identities for the multivariate normal with non-zero mean". In: *Hacettepe Journal of Mathematics and Statistics* 43.4 (2014), pp. 679–687 (cit. on pp. 5, 96).

[YCS14] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. "Alternating Minimization for Mixed Linear Regression". In: *Proceedings of the 31st International Conference on Machine Learning*. Vol. 32. Proceedings of Machine Learning Research 2. PMLR, 2014, pp. 613–621 (cit. on p. 14).

[YWS15] Yi Yu, Tengyao Wang, and Richard J. Samworth. "A useful variant of the Davis–Kahan theorem for statisticians". In: *Biometrika* 102.2 (2015), pp. 315–323 (cit. on p. 87).

[ZJD16] Kai Zhong, Prateek Jain, and Inderjit S. Dhillon. "Mixed Linear Regression with Multiple Components". In: *Advances in Neural Information Processing Systems*. Vol. 29. Curran Associates, Inc., 2016 (cit. on p. 14).

[ZSJB+17] Kai Zhong, Zhao Song, Prateek Jain, Peter L. Bartlett, and Inderjit S. Dhillon. "Recovery Guarantees for One-Hidden-Layer Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 4140–4149 (cit. on p. 14).

[ZZ25] Manolis Zampetakis and Felix Zhou. "Private Statistical Estimation via Truncation". In: *Advances in Neural Information Processing Systems*. Vol. 38. 2025, pp. 99980–100014 (cit. on p. 14).

A Additional Preliminaries

A.1 Gaussian Moments and Elementary Bounds

Proposition A.1 (Moment-Generating Function of a Gaussian Quadratic Form; [RS08, Theorem 5.2b]). *Let $G_0 \sim N(0,1)$, $m \in \mathbb{R}$, and $\tau \geq 0$. For every $a \in \mathbb{R}$ with $2a\tau^2 < 1$,*

$$\mathbb{E}_{G_0}\left[e^{a(m+\tau G_0)^2}\right] = \frac{\exp\left(\dfrac{am^2}{1-2a\tau^2}\right)}{\sqrt{1-2a\tau^2}}.$$

The following is a Gaussian fourth-moment specialization of the standard bounds for sub-Gaussian maxima; see [Ver18, Proposition 2.5.2 and Exercise 2.5.10]. The proof verifies the numerical constant.

Proposition A.2 (Fourth Moment of a Gaussian Maximum). *Let $\boldsymbol{G} = (G_1, \dots, G_n)$ be a Gaussian vector, where $n \geq 2$. Then*

$$\left(\mathbb{E}_{\boldsymbol{G}}\left[\max_{j\in[n]} |G_j|^4\right]\right)^{1/4} \leq \max_{j\in[n]} |\mathbb{E}_{\boldsymbol{G}}[G_j]| + 3\sqrt{\log n}\left(\max_{j\in[n]} \mathrm{Var}_{\boldsymbol{G}}(G_j)\right)^{1/2}.$$

Proof. We first bound the maximum after subtracting the means. Put $p = 6\log n > 4$. Each centered variable $G_j - \mathbb{E}_{\boldsymbol{G}}[G_j]$ has the law of a linear polynomial in a standard Gaussian, with L^2 norm $\mathrm{Var}_{\boldsymbol{G}}(G_j)^{1/2}$. Gaussian hypercontractivity (Proposition 4.4) therefore gives

$$\begin{aligned}
&\left(\mathbb{E}_{\boldsymbol{G}} \max_{j\in[n]} |G_j - \mathbb{E}_{\boldsymbol{G}}[G_j]|^4\right)^{1/4} \\
&\quad\leq \left(\mathbb{E}_{\boldsymbol{G}} \max_{j\in[n]} |G_j - \mathbb{E}_{\boldsymbol{G}}[G_j]|^p\right)^{1/p} && (p > 4) \\
&\quad\leq \left(\sum_{j\in[n]} \mathbb{E}_{\boldsymbol{G}} |G_j - \mathbb{E}_{\boldsymbol{G}}[G_j]|^p\right)^{1/p} \\
&\quad\leq n^{1/p}\sqrt{p-1}\left(\max_{j\in[n]} \mathrm{Var}_{\boldsymbol{G}}(G_j)\right)^{1/2} && (\text{by Proposition 4.4}) \\
&\quad\leq e^{1/6}\sqrt{6\log n}\left(\max_{j\in[n]} \mathrm{Var}_{\boldsymbol{G}}(G_j)\right)^{1/2} && (n^{1/p} = e^{1/6};\ p = 6\log n) \\
&\quad\leq 3\sqrt{\log n}\left(\max_{j\in[n]} \mathrm{Var}_{\boldsymbol{G}}(G_j)\right)^{1/2}.
\end{aligned}$$

Applying the triangle inequality inside the maximum and then Minkowski's inequality gives

$$\left(\mathbb{E}_{\boldsymbol{G}} \max_{j\in[n]} |G_j|^4\right)^{1/4} \leq \max_{j\in[n]} |\mathbb{E}_{\boldsymbol{G}}[G_j]| + \left(\mathbb{E}_{\boldsymbol{G}} \max_{j\in[n]} |G_j - \mathbb{E}_{\boldsymbol{G}}[G_j]|^4\right)^{1/4}.$$

Combining these bounds proves the claim. □

Proposition A.3 (Truncated Gaussian Second Moment; cf. [Han22, Theorem 5.8]). *Let G be a nondegenerate Gaussian with $\mathbb{E}_G[G] \leq 0$. Then*

$$\mathbb{E}_G[G^2 \mathbf{1}_{\{G>0\}}] \leq \mathrm{Var}_G(G)\,\mathbb{P}_G(G > 0).$$

Proof. Let ϕ_G denote the density of G. The Gaussian density satisfies

$$\mathrm{Var}_G(G)\phi_G'(t) = (\mathbb{E}_G[G] - t)\phi_G(t).$$

Hence

$$\begin{aligned}
\mathbb{E}_G[G^2\mathbf{1}_{\{G>0\}}] &= \int_0^\infty t^2\phi_G(t)\,\mathrm{d}t \\
&= \mathbb{E}_G[G]\int_0^\infty t\phi_G(t)\,\mathrm{d}t - \mathrm{Var}_G(G)\int_0^\infty t\phi_G'(t)\,\mathrm{d}t \\
&= \mathbb{E}_G[G]\,\mathbb{E}_G[G\mathbf{1}_{\{G>0\}}] + \mathrm{Var}_G(G)\int_0^\infty \phi_G(t)\,\mathrm{d}t \\
&\qquad\qquad \text{(integration by parts; } [t\phi_G(t)]_0^\infty = 0\text{)} \\
&= \mathbb{E}_G[G]\,\mathbb{E}_G[G\mathbf{1}_{\{G>0\}}] + \mathrm{Var}_G(G)\mathbb{P}_G(G>0) \\
&\le \mathrm{Var}_G(G)\mathbb{P}_G(G>0). \qquad\qquad (\mathbb{E}_G[G] \le 0;\ G\mathbf{1}_{\{G>0\}} \ge 0)
\end{aligned}$$

□

A.2 Linear Algebra

Proposition A.4 (Gram-Matrix Identities; Follows from [Bha97, Section I.2])**.** *For every real matrix B, the matrices $B^\top B$ and $BB^\top$ have the same nonzero eigenvalues, with multiplicity. Moreover,*

$$\left\|B^\top B\right\|_{\mathrm{op}} = \left\|BB^\top\right\|_{\mathrm{op}} = \|B\|_{\mathrm{op}}^2 .$$

B Deferred Proofs for Section 4

B.1 Hermite Polynomials

Proof of Proposition 4.5. We first verify orthogonality for products of normalized Hermite polynomials over independent Gaussian coordinates. Let $\boldsymbol{G} = (G_1, \ldots, G_n) \sim N(\mathbf{0}, I_n)$, and let $a_1, \ldots, a_n, b_1, \ldots, b_n$ be nonnegative integers. Then

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{G}}\left[\left(\prod_{j=1}^n \frac{\mathrm{He}_{a_j}(G_j)}{\sqrt{a_j!}}\right)\left(\prod_{j=1}^n \frac{\mathrm{He}_{b_j}(G_j)}{\sqrt{b_j!}}\right)\right] \\
&= \prod_{j=1}^n \mathbb{E}_{G_j}\left[\frac{\mathrm{He}_{a_j}(G_j)\,\mathrm{He}_{b_j}(G_j)}{\sqrt{a_j!b_j!}}\right] \qquad (123) \\
&= \begin{cases} 1, & a_j = b_j \text{ for every } j, \\ 0, & \text{otherwise.} \end{cases}
\end{aligned}$$

The first equality uses independence. The second applies Hermite orthogonality (Section 4.4) in each coordinate.

For a symmetric tensor A, use the coordinate counts $\boldsymbol{\alpha}$ from Section 4.4. The complex Gaussian

definition and independence of the coordinates of $\boldsymbol{X}'$ give, for indices $j_1, \dots, j_d$ with counts $\boldsymbol{\alpha}$,

$$\begin{aligned}
[\mathsf{H}_d(\boldsymbol{x})]_{j_1,\dots,j_d} &= \frac{1}{\sqrt{d!}} \operatorname{Re}\left(\mathbb{E}_{\boldsymbol{X}'} \left[\prod_{\ell=1}^{d} (x_{j_\ell} + \mathrm{i} X'_{j_\ell}) \right] \right) \\
&= \frac{1}{\sqrt{d!}} \operatorname{Re}\left(\prod_{j=1}^{n} \mathbb{E}_{X'_j}[(x_j + \mathrm{i} X'_j)^{\alpha_j}] \right) && \text{(independence)} \\
&= \frac{1}{\sqrt{d!}} \prod_{j=1}^{n} \mathrm{He}_{\alpha_j}(x_j). && \text{(definition of } \mathrm{He}_{\alpha_j}\text{)}
\end{aligned}$$

The last product is real.

Grouping the tensor entries gives the expansion in Proposition 4.5. Product orthogonality then gives the stated isometry. □

B.2 Heavy-Tail Mean Estimation

Proof of Lemma 4.7. Choose

$$\tau = \left\lceil 3 \log\left(\frac{p}{\delta}\right) \right\rceil, \qquad q = \left\lceil \frac{1024 \mathsf{v} \tau}{\varepsilon^2} \right\rceil,$$

and draw q independent copies of $\boldsymbol{W}$. For each $j \in [p]$, let $W_{j,(1)} \le \cdots \le W_{j,(q)}$ be the order statistics of the j-th coordinates, and return

$$\widehat{\mu}_j = \frac{1}{q - 2\tau} \sum_{t=\tau+1}^{q-\tau} W_{j,(t)}.$$

Here $q \ge 1024\tau$, since $\varepsilon^2 \le \mathsf{v}$. All coordinates use the same observations.

Apply [OOR25, Theorem 1.1.1] to each coordinate with $x = \sqrt{2\tau}$. Since $\tau \ge 1$,

$$(\sqrt{2}+1)^2(2\tau+2) \le 24\tau \le q,$$

which verifies its sample-size condition. The coordinate variances are at most v, so its error bound is at most

$$\begin{aligned}
\frac{(3\sqrt{2} + 8 + (4+4\sqrt{2})\sqrt{2\tau})\sqrt{\mathsf{v}}}{\sqrt{q}} &\le (7\sqrt{2}+16)\sqrt{\frac{\mathsf{v}\tau}{q}} \\
&\le 32\sqrt{\frac{\mathsf{v}\tau}{q}} \le \varepsilon. && \text{(choice of } q\text{)}
\end{aligned}$$

A union bound gives failure probability at most

$$4pe^{-\tau} \le \frac{4\delta^3}{p^2} \le \delta.$$

Since $\mathsf{v}/\varepsilon^2 \ge 1$ and $\log(p/\delta) > 1$, the ceilings change the sample count by at most an absolute factor. The trimmed sum can be computed in $O(q)$ operations per coordinate by finding the two cutoff order statistics and summing the middle values. □

C Deferred Proofs for Section 5

C.1 Weights and Scaled Regressors

We specialize the norm and separation bounds of Theorem 5.1(iv) to the hyperparameter choice used in ImplicitTensorDecomposition.

Proposition C.1 (Geometry of the Scaled Regressors). *Under the hypotheses and notation of Theorem 5.1, the scaled regressors satisfy*

$$K^{-3} \le \|\boldsymbol{a}_i\|_2 \le K^3 \quad (i \in [k]), \qquad \|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2 \ge K^{-3} \quad (i \ne j). \tag{124}$$

Proof. By the choice of Hyperparameters 4.1,

$$K \ge \frac{128B^2}{\Delta^2} \ge \max\left\{3B, \frac{\sqrt{5}(1+B^2)}{\Delta}\right\}, \tag{125}$$

$$2s+1 \le \frac{5T^2}{4K^4}, \tag{126}$$

$$2s+1 \ge \frac{T^2}{4K^4}, \tag{127}$$

$$T+1 \le \frac{3T}{2}. \tag{128}$$

The first inequality follows from $(1+B^2)^3 \ge 8B^2$ and $\Delta \le 1$; the remaining bounds use $T^2 = 4K^4 s$ and $s \ge 1$. The scale bounds in Theorem 5.1(iv) give

$$\begin{aligned}
\|\boldsymbol{a}_i\|_2 &\ge \frac{T\Delta}{2(1+B^2)\sqrt{2s+1}} && \text{(by Theorem 5.1(iv) and Assumption 1.2)}\\
&\ge \frac{K^2\Delta}{\sqrt{5}(1+B^2)} && \text{(choice of } T, s \text{ (126))}\\
&\ge K, && \text{(choice of } K \text{ (125))}\\
\|\boldsymbol{a}_i\|_2 &\le \frac{(T+1)B}{\sqrt{2s+1}} && \text{(by Theorem 5.1(iv) and Assumption 1.2)}\\
&\le 3K^2B && \text{(choice of } T, s \text{ (128), (127))}\\
&\le K^3. && \text{(choice of } K \text{ (125))}
\end{aligned}$$

For distinct i, j, the separation bound in Theorem 5.1(iv) now gives

$$\begin{aligned}
\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2^2 &\ge K^{-1}\max\{\|\boldsymbol{a}_i\|_2^2, \|\boldsymbol{a}_j\|_2^2\} && \text{(by Theorem 5.1(iv))}\\
&\ge K \ge K^{-6}. && (\|\boldsymbol{a}_i\|_2 \ge K \text{ and } K \ge 1)
\end{aligned}$$

Taking square roots, and using $K \ge K^{-3}$ for the norm lower bound, proves (124). □

C.2 Single-Response Moments

We prove the conditional Hermite identity in Proposition 5.2 by applying the scalar Gaussian noise identity below to projections of the covariate.

Proposition C.2 (Gaussian Noise Operator on Hermite Polynomials; [ODo14, Definition 11.12 and Proposition 11.37]). *For a polynomial $f : \mathbb{R} \to \mathbb{R}$ and $c \in [-1, 1]$, the Gaussian noise operator is given by $(U_c f)(x) = \mathbb{E}_g[f(cx + \sqrt{1-c^2}\, g)]$, where $g \sim N(0,1)$. For every integer $d \geq 0$, it satisfies $U_c \operatorname{He}_d = c^d \operatorname{He}_d$. Thus, if U, V are jointly standard Gaussian with correlation c, then for every $x \in \mathbb{R}$,*

$$\mathbb{E}_{U,V}[\operatorname{He}_d(U) \mid V = x] = c^d \operatorname{He}_d(x). \tag{129}$$

Proposition 5.2 (Conditional Hermite Identity). *Let $\boldsymbol{G} \in \mathbb{R}^n$ and $Y \in \mathbb{R}$ be centered jointly Gaussian, with $\boldsymbol{G} \sim N(\mathbf{0}, I_n)$ and $\operatorname{Var}(Y) = \sigma^2 > 0$. Put $\boldsymbol{v} = \mathbb{E}_{\boldsymbol{G},Y}[\boldsymbol{G}Y]$. For every integer $d \geq 0$ and $t \in \mathbb{R}$,*

$$\mathbb{E}_{\boldsymbol{G},Y}[\mathsf{H}_d(\boldsymbol{G}) \mid Y = t] = \frac{\operatorname{He}_d(t/\sigma)}{\sigma^d \sqrt{d!}} \boldsymbol{v}^{\otimes d}.$$

Proof. We apply the Gaussian noise identity (Proposition C.2) to scalar projections of $\boldsymbol{G}$.

The case $d = 0$ is immediate. Fix $d \geq 1$ and $\boldsymbol{\theta} \neq 0$. The complex Gaussian definition of the Hermite tensor (Section 4.4) gives, for every $\boldsymbol{x} \in \mathbb{R}^n$,

$$\begin{aligned}
\left\langle \mathsf{H}_d(\boldsymbol{x}), \boldsymbol{\theta}^{\otimes d} \right\rangle &= \frac{1}{\sqrt{d!}} \mathbb{E}_{\boldsymbol{X}'}[\operatorname{Re}(\langle \boldsymbol{\theta}, \boldsymbol{x} + \mathrm{i}\boldsymbol{X}' \rangle^d)] \\
&= \frac{\|\boldsymbol{\theta}\|_2^d}{\sqrt{d!}} \operatorname{He}_d\left(\frac{\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle}{\|\boldsymbol{\theta}\|_2} \right). && (\langle \boldsymbol{\theta}, \boldsymbol{X}' \rangle / \|\boldsymbol{\theta}\|_2 \sim N(0,1))
\end{aligned}$$

The variables $\langle \boldsymbol{\theta}, \boldsymbol{G} \rangle / \|\boldsymbol{\theta}\|_2$ and Y/σ are standard Gaussians with correlation $\langle \boldsymbol{\theta}, \boldsymbol{v} \rangle / (\sigma \|\boldsymbol{\theta}\|_2)$. Thus

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{G},Y}\left[\left\langle \mathsf{H}_d(\boldsymbol{G}), \boldsymbol{\theta}^{\otimes d} \right\rangle \mid Y = t \right] &= \frac{\|\boldsymbol{\theta}\|_2^d}{\sqrt{d!}} \mathbb{E}_{\boldsymbol{G},Y}\left[\operatorname{He}_d\left(\frac{\langle \boldsymbol{\theta}, \boldsymbol{G} \rangle}{\|\boldsymbol{\theta}\|_2} \right) \,\middle|\, Y = t \right] \\
&= \frac{\|\boldsymbol{\theta}\|_2^d}{\sqrt{d!}} \left(\frac{\langle \boldsymbol{\theta}, \boldsymbol{v} \rangle}{\sigma \|\boldsymbol{\theta}\|_2} \right)^d \operatorname{He}_d(t/\sigma) && \text{(by (129))} \\
&= \frac{\operatorname{He}_d(t/\sigma)}{\sigma^d \sqrt{d!}} \langle \boldsymbol{\theta}, \boldsymbol{v} \rangle^d .
\end{aligned}$$

Both tensors are symmetric, and the tensors $\boldsymbol{\theta}^{\otimes d}$ span the symmetric order-d tensor space. This proves the identity. □

C.3 Moment Decomposition Error

We prove the conditional second-moment bound used in Section 5.1.2. Conditioning changes only one Gaussian coordinate. Orthogonality in the other coordinates reduces the bound to scalar moments, which we control by the complex Gaussian definition and Jensen's inequality.

Proposition 5.5 (Conditional Hermite Second Moment). *Let $\boldsymbol{G} \in \mathbb{R}^n$ and $Y \in \mathbb{R}$ be centered jointly Gaussian, with $\boldsymbol{G} \sim N(\mathbf{0}, I_n)$ and $\operatorname{Var}(Y) = \sigma^2 > 0$. Put $\boldsymbol{v} = \mathbb{E}_{\boldsymbol{G},Y}[\boldsymbol{G}Y]$. For every integer $d \geq 0$, symmetric order-d tensor A, $t \in \mathbb{R}$, and $0 < \rho < 1/2$,*

$$\left(\mathbb{E}_{\boldsymbol{G},Y}[\langle \mathsf{H}_d(\boldsymbol{G}), \mathsf{A} \rangle^2 \mid Y = t] \right)^{1/2} \leq \frac{\rho^{-d/2}}{\sqrt{1-2\rho}} \exp\left(\frac{\rho \|\boldsymbol{v}\|_2^2 t^2}{2\sigma^4(1-2\rho)} \right) \|\mathsf{A}\|_F .$$

Proof. Choose orthonormal coordinates in which $\boldsymbol{v} = \|\boldsymbol{v}\|_2 \, \boldsymbol{e}_1$; if $\boldsymbol{v} = 0$, any orthonormal coordinates suffice. Express $\boldsymbol{G}$ and A in these coordinates. Rotational invariance of the auxiliary Gaussian in the definition of H_d shows that the tensor pairing and Frobenius norm are unchanged. The conditional Gaussian formula (Proposition 4.3) gives

$$\begin{aligned} \boldsymbol{G} \mid Y = t &\sim N\big(\mu \boldsymbol{e}_1, \mathrm{Diag}(\tau^2, 1, \dots, 1)\big)\,, \\ \mu &= \frac{\|\boldsymbol{v}\|_2 \, t}{\sigma^2}, \\ \tau^2 &= 1 - \frac{\|\boldsymbol{v}\|_2^2}{\sigma^2} \in [0, 1]. \end{aligned}$$

Under this conditional law, write $G_1 = \mu + \tau G_0$, where $G_0 \sim N(0,1)$. The remaining coordinates $G_2, \dots, G_n$ are independent standard Gaussians, independent of G_0.

We first use this independence to reduce the tensor second moment to scalar moments. Squaring the expansion in Proposition 4.5 gives the following double sum. Orthogonality forces $\alpha_j = \beta_j$ for $j \geq 2$; the common total degree then gives $\alpha_1 = \beta_1$.

$$\begin{aligned} \mathbb{E}_{\boldsymbol{G},Y}[\langle \mathsf{H}_d(\boldsymbol{G}), \mathsf{A}\rangle^2 \mid Y = t] &= \sum_{\substack{|\boldsymbol{\alpha}|=d \\ |\boldsymbol{\beta}|=d}} \frac{d!\, A_{\boldsymbol{\alpha}} A_{\boldsymbol{\beta}}}{\sqrt{\boldsymbol{\alpha}!\boldsymbol{\beta}!}} \, \mathbb{E}_{G_0}\left[\frac{\mathrm{He}_{\alpha_1}(\mu + \tau G_0)\, \mathrm{He}_{\beta_1}(\mu + \tau G_0)}{\sqrt{\alpha_1!\beta_1!}}\right] \\ &\qquad \times \prod_{j=2}^{n} \mathbb{E}_{G_j}\left[\frac{\mathrm{He}_{\alpha_j}(G_j)\, \mathrm{He}_{\beta_j}(G_j)}{\sqrt{\alpha_j!\beta_j!}}\right] \\ &= \sum_{|\boldsymbol{\alpha}|=d} \frac{d!}{\boldsymbol{\alpha}!} A_{\boldsymbol{\alpha}}^2 \mathbb{E}_{G_0}\left[\frac{\mathrm{He}_{\alpha_1}(\mu + \tau G_0)^2}{\alpha_1!}\right] \\ &\qquad\qquad \text{(Hermite orthogonality (14); } |\boldsymbol{\alpha}| = |\boldsymbol{\beta}| = d) \\ &\leq \|\mathsf{A}\|_F^2 \max_{0 \leq \ell \leq d} \frac{\mathbb{E}_{G_0}[\mathrm{He}_\ell(\mu + \tau G_0)^2]}{\ell!}. \qquad \text{(by (15))} \quad (130) \end{aligned}$$

It remains to bound these scalar moments. Let $g \sim N(0,1)$ be independent of G_0. We use the complex Gaussian definition of He_ℓ and Jensen's inequality, followed by the Gaussian quadratic-form moment-generating formula (Proposition A.1). Since $\tau^2 \leq 1$ and $0 < \rho < 1/2$, both exponential moments below are finite. For $0 \leq \ell \leq d$,

$$\begin{aligned} \frac{\mathbb{E}_{G_0}[\mathrm{He}_\ell(\mu + \tau G_0)^2]}{\ell!} &= \frac{1}{\ell!} \mathbb{E}_{G_0}\left[\left|\mathbb{E}_g[(\mu + \tau G_0 + \mathrm{i}g)^\ell]\right|^2\right] && \text{(definition of } \mathrm{He}_\ell) \\ &\leq \frac{\mathbb{E}_{G_0,g}[((\mu + \tau G_0)^2 + g^2)^\ell]}{\ell!} && \text{(Jensen's inequality)} \\ &\leq \rho^{-\ell} \mathbb{E}_{G_0,g}\left[e^{\rho((\mu+\tau G_0)^2 + g^2)}\right] && (x^\ell/\ell! \leq \rho^{-\ell} e^{\rho x} \text{ for } x \geq 0) \\ &= \rho^{-\ell} \mathbb{E}_{G_0}[e^{\rho(\mu+\tau G_0)^2}] \mathbb{E}_g[e^{\rho g^2}] && \text{(independence)} \\ &= \frac{\rho^{-\ell} \exp\left(\frac{\rho\mu^2}{1-2\rho\tau^2}\right)}{\sqrt{(1-2\rho\tau^2)(1-2\rho)}} && \text{(by Proposition A.1)} \\ &\leq \frac{\rho^{-\ell}}{1-2\rho} \exp\left(\frac{\rho\mu^2}{1-2\rho}\right). && (\tau^2 \leq 1) \quad (131) \end{aligned}$$

Combining (130) and (131) gives

$$\left(\mathbb{E}_{G,Y}[\langle \mathsf{H}_d(G), \mathsf{A}\rangle^2 \mid Y = t]\right)^{1/2} \le \frac{\rho^{-d/2}}{\sqrt{1-2\rho}} \exp\left(\frac{\rho\mu^2}{2(1-2\rho)}\right) \|\mathsf{A}\|_F \qquad (\rho^{-\ell} \le \rho^{-d})$$

$$= \frac{\rho^{-d/2}}{\sqrt{1-2\rho}} \exp\left(\frac{\rho \|v\|_2^2 t^2}{2\sigma^4(1-2\rho)}\right) \|\mathsf{A}\|_F . \qquad (\text{definition of } \mu)$$

This proves the claim. □

D Deferred Proofs for Section 6

D.1 Iterative Projections

Proposition D.1 (Hyperparameter Bounds for Iterative Projections)**.** *Use the hyperparameters from Hyperparameters 4.1 and put* $\varepsilon_{\text{est}} = e^{-K^6 s}$*. Then*

$$e^{T^2}\left(1 + \sqrt{\frac{k}{2\pi}}\right)\varepsilon_{\text{est}} + \left(kK^{4s} + K^{8s}\right) e^{-T^2/(2K)} \le e^{-K^2 s}, \tag{132}$$

$$\frac{n^4 + sn^2k^2}{\varepsilon_{\text{est}}^2} 100^s \le (nk)^{4K^{10}+12K^4}. \tag{133}$$

Proof. By the choice of Hyperparameters 4.1,

$$K \ge 128, \tag{134}$$

$$s \ge K^4 \log n \ge 7, \tag{135}$$

$$s \le 2K^4 \log n, \tag{136}$$

$$T^2 = 4K^4 s, \tag{137}$$

$$\varepsilon_{\text{est}} = e^{-K^6 s}. \tag{138}$$

Error bound. We bound each of the four terms on the left of (132). The choice of s and $k \le n$ give $k \le e^{s/K^4} \le K^s$ and $\sqrt{k} \le e^s$. For $K \ge 128$,

$$K^6 - 4K^4 - 1 \ge (K^2 - 5)K^4 \ge 3K^2, \tag{139}$$

$$2K^3 - 8\log K \ge 2K^3 - 8K \ge 3K^2. \qquad (\log K \le K) \tag{140}$$

Consequently,

$$\begin{aligned}
e^{T^2}\varepsilon_{\mathrm{est}} &= e^{-(K^6-4K^4)s} && \text{(choice of } T, \varepsilon_{\mathrm{est}} \text{ (137), (138))}\\
&\le e^{-3K^2 s}, && \text{(by (139))}\\
e^{T^2}\sqrt{\frac{k}{2\pi}}\,\varepsilon_{\mathrm{est}} &\le e^{-(K^6-4K^4-1)s} && (\sqrt{k} \le e^s,\ (2\pi)^{-1/2} \le 1)\\
&\le e^{-3K^2 s}, && \text{(by (139))}\\
e^{8s\log K - T^2/(2K)} &= e^{-(2K^3-8\log K)s} && \text{(choice of } T \text{ (137))}\\
&\le e^{-3K^2 s}, && \text{(by (140))}\\
ke^{4s\log K - T^2/(2K)} &\le e^{-(2K^3-5\log K)s} && (k \le K^s, \text{ choice of } T \text{ (137))}\\
&\le e^{-3K^2 s}. && \text{(by (140))}
\end{aligned}$$

Their sum is at most $4e^{-3K^2 s} \le e^{-K^2 s}$, which proves (132).

Sample factor. We now bound the expression in (133). The choice of s gives $s+1 \le 2K^4 n$, so

$$\begin{aligned}
\frac{1}{\varepsilon_{\mathrm{est}}^2} &= e^{2K^6 s} && \text{(choice of } \varepsilon_{\mathrm{est}} \text{ (138))}\\
&\le (nk)^{4K^{10}}, && \text{(choice of } s \text{ (136))}\\
100^s(n^4 + sn^2k^2) &\le (s+1)(nk)^{10K^4+4} && \text{(choice of } s \text{ (136)}, \log 100 \le 5)\\
&\le 2K^4(nk)^{10K^4+5} && (s+1 \le 2K^4 n,\ n \le nk)\\
&\le (nk)^{12K^4}. && (2K^4 \le (nk)^{K^4},\ K^4 \ge 5)
\end{aligned}$$

Multiplying these bounds proves (133). □

D.2 Observable-Operator Learning

We prove the largest-gap bound used to choose the block, then record the polar-factor and Sylvester bounds used in the truncation and approximate-orthonormality calculations. We prove Lemma D.4 and then record the spherical anti-concentration and eigenvector perturbation bounds used in Section 6.3.

Proof of Proposition 6.8. We first compare the estimated eigenvalues with those of Λ. Pad the clipped estimated eigenvalues and the eigenvalues of $WW^\top$ with zeros through index $q+1$. The spectrum identity (Proposition A.4) identifies the padded eigenvalues of $WW^\top$ with those of $W^\top W$, also padded by a zero at index $q+1$. Clipping at zero cannot increase the error relative to these nonnegative eigenvalues. Weyl's inequality (Proposition 4.6) therefore gives

$$\begin{aligned}
\left|\widehat{\lambda}_h - \lambda_h(\Lambda)\right| &\le \left\|A - WW^\top\right\|_{\mathrm{op}} + \left\|W^\top W - \Lambda\right\|_{\mathrm{op}}\\
&\le \frac{\|\Lambda\|_{\mathrm{op}}}{8q} + \frac{\|\Lambda\|_{\mathrm{op}}}{8q} && \text{(assumed perturbation bounds)}\\
&= \frac{\|\Lambda\|_{\mathrm{op}}}{4q} \qquad (1 \le h \le q). && (141)
\end{aligned}$$

The same bound holds at $h = q+1$, where both eigenvalues are zero. In particular, $\widehat{\lambda}_1 \geq 3\|\Lambda\|_{\mathrm{op}}/4$.

We now use this comparison to bound the upper endpoint of the largest gap and its separation from the next true eigenvalue. Telescoping gives

$$\begin{aligned}
\widehat{\lambda}_{i_\star} - \widehat{\lambda}_{i_\star+1} &\geq \frac{1}{p}\sum_{h=1}^{p}(\widehat{\lambda}_h - \widehat{\lambda}_{h+1}) && \text{(choice of } i_\star\text{)} \\
&= \frac{\widehat{\lambda}_1}{p} \geq \frac{3\|\Lambda\|_{\mathrm{op}}}{4q}, && (p \leq q) \\
\widehat{\lambda}_{i_\star} &\geq \widehat{\lambda}_{i_\star} - \widehat{\lambda}_{i_\star+1} \geq \frac{\|\Lambda\|_{\mathrm{op}}}{2q}, \\
\widehat{\lambda}_{i_\star} - \lambda_{i_\star+1}(\Lambda) &\geq \widehat{\lambda}_{i_\star} - \widehat{\lambda}_{i_\star+1} - \frac{\|\Lambda\|_{\mathrm{op}}}{4q} && \text{(by (141))} \\
&\geq \frac{\|\Lambda\|_{\mathrm{op}}}{2q}.
\end{aligned}$$

□

Proposition D.2 (Polar-Factor Bound). *Let $B \in \mathbb{R}^{p\times p}$, and let Z be an orthogonal polar factor of B. Writing $\sigma_h(B)$ for the singular values, we have*

$$\|B - Z\|_F^2 = \sum_{h=1}^{p}(\sigma_h(B) - 1)^2 \leq \left\|BB^\top - I\right\|_F^2.$$

The polar factor and its Frobenius-distance formula are given in [Hig86, Theorem 1.1 and Corollary 2.3]; the inequality uses $|\sigma - 1| \leq |\sigma^2 - 1|$ for $\sigma \geq 0$.

Proposition D.3 (Sylvester's Equation; [Bha97, Theorem VII.2.8]). *Let $A \in \mathbb{R}^{p\times p}$ and $B \in \mathbb{R}^{q\times q}$ be symmetric. Suppose $|\lambda_h(A) - \lambda_i(B)| \geq g > 0$ for every $h \in [p]$ and $i \in [q]$. If $AX - XB = C$, then*

$$\|X\|_F \leq \frac{\|C\|_F}{g}.$$

Lemma D.4 (Spectral Truncation and Approximate Orthonormality). *Under the hypotheses of Theorem 6.5, let $F = [\boldsymbol{f}_i]_{i\in Q}$ be the rescaled component matrix from (65), and let J index the $i_\star$ largest component weights in the fixed order.*

(i) The components outside J satisfy

$$\left\|F_{Q\setminus J}\right\|_F^2 \leq |Q|^3\left(\varepsilon_{\mathrm{mom}} + \varepsilon_{\mathrm{proj}} + \varepsilon_\perp\right).$$

(ii) For every $\boldsymbol{b} \in \mathbb{R}^n$, the contribution of components outside J to the rescaled odd moment satisfies

$$\left\|\sum_{i\in Q\setminus J}\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle\ \boldsymbol{f}_i\boldsymbol{f}_i^\top\right\|_F \leq K^3\left\|F_{Q\setminus J}\right\|_F^2\|\boldsymbol{b}\|_2.$$

(iii) Let $Z = [z_i]_{i\in J}$ be an orthogonal polar factor of F_J. For every $b \in \mathbb{R}^n$, replacing the vectors f_i in J by z_i gives the error bound

$$\left\| \sum_{i\in J} \langle b, a_i \rangle \left(f_i f_i^\top - z_i z_i^\top \right) \right\|_F \leq 3K^3 \left(2|Q|^2 \varepsilon_{\mathrm{mom}} + \left\| F_{Q\setminus J} \right\|_F^2 \right) \|b\|_2 .$$

Proof. We first recall the perturbation estimates used to justify rescaling. We then bound the components outside J using Sylvester's equation and use their bound to control the correction to an orthonormal basis.

Let

$$H = \left[\sqrt{\lambda_i^{(2s)}}\, u_i^{\otimes s} \right]_{i\in Q}, \qquad \Lambda = \mathrm{diag}_{i\in Q}(\lambda_i^{(2s)}),$$

with components ordered by decreasing weight. Then $A_Q^{(s)} = HH^\top$ and $\mathrm{tr}(A_Q^{(s)}) = \mathrm{tr}(\Lambda) \leq |Q|\, \|\Lambda\|_{\mathrm{op}}$. The even-moment, projection, and approximate-orthogonality assumptions give the perturbation bounds established in the proof of Theorem 6.5. We recall them here in the form used below.

$$\left\| \widehat{A} - C_s A_Q^{(s)} C_s^\top \right\|_F \leq |Q| \varepsilon_{\mathrm{mom}} \|\Lambda\|_{\mathrm{op}}, \tag{142}$$

$$\left\| H^\top C_s^\top C_s H - \Lambda \right\|_{\mathrm{op}} \leq \left(|Q| \varepsilon_{\mathrm{proj}} + \varepsilon_\perp \right) \|\Lambda\|_{\mathrm{op}} . \tag{143}$$

Since

$$|Q|\varepsilon_{\mathrm{mom}} + |Q|\varepsilon_{\mathrm{proj}} + \varepsilon_\perp \leq |Q|\left(\varepsilon_{\mathrm{mom}} + \varepsilon_{\mathrm{proj}} + \varepsilon_\perp \right) \leq \frac{1}{16|Q|},$$

Proposition 6.8, with $W = C_s H$, gives

$$\widehat{\lambda}_{i_\star} \geq \frac{\|\Lambda\|_{\mathrm{op}}}{2|Q|}, \tag{144}$$

$$\widehat{\lambda}_{i_\star} - \lambda_{i_\star+1}(\Lambda) \geq \frac{\|\Lambda\|_{\mathrm{op}}}{2|Q|}. \tag{145}$$

Thus $\widehat{\Lambda}$ is positive definite. We use $F = \widehat{\Lambda}^{-1/2} \widehat{V} C_s H$ throughout.

Since $\widehat{V}\widehat{A}\widehat{V}^\top = \widehat{\Lambda}$, the even-moment estimation assumption (Definition 6.2(i)) gives

$$\begin{aligned}
\left\| FF^\top - I \right\|_F &= \left\| \widehat{\Lambda}^{-1/2} \widehat{V} (C_s A_Q^{(s)} C_s^\top - \widehat{A}) \widehat{V}^\top \widehat{\Lambda}^{-1/2} \right\|_F \\
&\leq \left\| \widehat{\Lambda}^{-1/2} \right\|_{\mathrm{op}}^2 \left\| \widehat{V} \right\|_{\mathrm{op}}^2 \left\| C_s A_Q^{(s)} C_s^\top - \widehat{A} \right\|_F && \text{(Frobenius product inequality)} \\
&= \frac{\left\| C_s A_Q^{(s)} C_s^\top - \widehat{A} \right\|_F}{\widehat{\lambda}_{i_\star}} \\
&\leq \frac{\varepsilon_{\mathrm{mom}}\, \mathrm{tr}(A_Q^{(s)})}{\widehat{\lambda}_{i_\star}} && \text{(by Definition 6.2(i))} \\
&\leq 2|Q|^2 \varepsilon_{\mathrm{mom}} && \text{(by (144))} \\
&\leq \frac{1}{8}. && \text{(by Assumption 6.5(e), } |Q| \leq k) \quad (146)
\end{aligned}$$

Consequently,

$$\begin{aligned}\|F\|_{\mathrm{op}} &= \sqrt{\|FF^\top\|_{\mathrm{op}}} \\ &\le \sqrt{1+\|FF^\top - I\|_F} \\ &\le \sqrt{\frac{9}{8}} < 2. \qquad \text{(by (146))} \end{aligned} \tag{147}$$

For the remaining factor, the Gram-matrix norm identity (Proposition A.4) and the projected-weight bound give

$$\begin{aligned}\|C_s H\|_{\mathrm{op}}^2 &= \left\|H^\top C_s^\top C_s H\right\|_{\mathrm{op}} \\ &\le \|\Lambda\|_{\mathrm{op}} + \left\|H^\top C_s^\top C_s H - \Lambda\right\|_{\mathrm{op}} \\ &\le 2\|\Lambda\|_{\mathrm{op}}, \qquad \text{(by (143) and Assumption 6.5(e))} \\ \left\|\widehat{\Lambda}^{-1/2}\right\|_{\mathrm{op}} \|C_s H\|_{\mathrm{op}} &\le \sqrt{\frac{2\|\Lambda\|_{\mathrm{op}}}{\widehat{\lambda}_{i_\star}}} \\ &\le 2\sqrt{|Q|}. \qquad \text{(by (144))} \end{aligned} \tag{148}$$

Components outside J. We now prove Lemma D.4(i). If $J = Q$, its left-hand side is zero. Otherwise, restricting the residual to the columns outside J gives the Sylvester equation

$$\widehat{\Lambda} F_{Q\setminus J} - F_{Q\setminus J}\,\mathrm{diag}_{i\in Q\setminus J}(\lambda_i^{(2s)}) = (\widehat{\Lambda}F - F\Lambda)_{Q\setminus J}.$$

Since J indexes the $i_\star$ largest weights, for $1 \le h \le i_\star$ and $i \in Q \setminus J$,

$$\begin{aligned}\widehat{\lambda}_h - \lambda_i^{(2s)} &\ge \widehat{\lambda}_{i_\star} - \lambda_{i_\star+1}(\Lambda) \\ &\ge \frac{\|\Lambda\|_{\mathrm{op}}}{2|Q|}. \qquad \text{(by (145))} \end{aligned} \tag{149}$$

Sylvester's equation (Proposition D.3) therefore gives

$$\begin{aligned}\|F_{Q\setminus J}\|_F &\le \frac{2|Q|}{\|\Lambda\|_{\mathrm{op}}} \left\|(\widehat{\Lambda}F - F\Lambda)_{Q\setminus J}\right\|_F \qquad \text{(by Proposition D.3)} \\ &\le \frac{2|Q|}{\|\Lambda\|_{\mathrm{op}}} \left\|\widehat{\Lambda}F - F\Lambda\right\|_F. \end{aligned} \tag{150}$$

We now bound the right-hand side of the Sylvester equation using the two perturbation errors. Using $\widehat{V}\widehat{A} = \widehat{\Lambda}\widehat{V}$ and $HH^\top = A_Q^{(s)}$, we obtain

$$\begin{aligned}\widehat{\Lambda}F - F\Lambda &= \left(\widehat{\Lambda}F - FH^\top C_s^\top C_s H\right) + F(H^\top C_s^\top C_s H - \Lambda) \\ &= \widehat{\Lambda}^{-1/2}\widehat{V}(\widehat{A} - C_s A_Q^{(s)} C_s^\top) C_s H + F(H^\top C_s^\top C_s H - \Lambda). \qquad \text{(by (65))} \end{aligned} \tag{151}$$

Since F has $|Q|$ columns,

$$\|F\|_F \le \sqrt{|Q|}\,\|F\|_{\mathrm{op}} \le 2\sqrt{|Q|}. \tag{152}$$

Using (151) and the Frobenius product inequality gives

$$\begin{aligned}
\left\|\widehat{\Lambda}F - F\Lambda\right\|_F &\le \left\|\widehat{\Lambda}^{-1/2}\right\|_{\mathrm{op}} \|C_s H\|_{\mathrm{op}} \left\|\widehat{A} - C_s A_Q^{(s)} C_s^\top\right\|_F \\
&\quad + \|F\|_F \left\|H^\top C_s^\top C_s H - \Lambda\right\|_{\mathrm{op}} && (\left\|\widehat{V}\right\|_{\mathrm{op}} = 1) \\
&\le 2\sqrt{|Q|}\Big(\left\|\widehat{A} - C_s A_Q^{(s)} C_s^\top\right\|_F \\
&\quad + \left\|H^\top C_s^\top C_s H - \Lambda\right\|_{\mathrm{op}}\Big) && \text{(by (148), (152))} \\
&\le 2|Q|^{3/2}\,\|\Lambda\|_{\mathrm{op}}\left(\varepsilon_{\mathrm{mom}} + \varepsilon_{\mathrm{proj}} + \varepsilon_\perp\right). && \text{(by (142), (143); } |Q| \ge 1) \quad (153)
\end{aligned}$$

Squaring (150) gives

$$\begin{aligned}
\|F_{Q\setminus J}\|_F^2 &\le \frac{4|Q|^2}{\|\Lambda\|_{\mathrm{op}}^2}\left\|\widehat{\Lambda}F - F\Lambda\right\|_F^2 && \text{(by (150))} \\
&\le 16|Q|^5\left(\varepsilon_{\mathrm{mom}} + \varepsilon_{\mathrm{proj}} + \varepsilon_\perp\right)^2 && \text{(by (153))} \\
&\le |Q|^3\left(\varepsilon_{\mathrm{mom}} + \varepsilon_{\mathrm{proj}} + \varepsilon_\perp\right). && \text{(by Assumption 6.5(e), } |Q| \le k) \quad (154)
\end{aligned}$$

This proves Lemma D.4(i), including the case $J = Q$.

Spectral truncation error. We prove Lemma D.4(ii) by bounding the weighted sum of outer products of the columns outside J. This sum equals $F_{Q\setminus J}\,\mathrm{diag}_{i\in Q\setminus J}(\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle)F_{Q\setminus J}^\top$, so

$$\begin{aligned}
&\left\|F_{Q\setminus J}\,\mathrm{diag}_{i\in Q\setminus J}(\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle)F_{Q\setminus J}^\top\right\|_F \\
&\quad \le \|F_{Q\setminus J}\|_F \left\|\mathrm{diag}_{i\in Q\setminus J}(\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle)\right\|_{\mathrm{op}} \|F_{Q\setminus J}\|_{\mathrm{op}} && \text{(Frobenius product inequality)} \\
&\quad \le K^3\,\|\boldsymbol{b}\|_2\,\|F_{Q\setminus J}\|_F^2. && \text{(Cauchy–Schwarz; } \|\boldsymbol{a}_i\|_2 \le K^3) \quad (155)
\end{aligned}$$

This proves Lemma D.4(ii).

Approximate orthonormality. We now prove Lemma D.4(iii). Choose an orthogonal polar factor $Z = [\boldsymbol{z}_i]_{i\in J}$ of the square matrix F_J. The orthogonal polar-factor bound (Proposition D.2) gives

$$\begin{aligned}
\|F_J - Z\|_F &\le \left\|F_J F_J^\top - I\right\|_F && \text{(by Proposition D.2)} \\
&= \left\|(FF^\top - I) - F_{Q\setminus J}F_{Q\setminus J}^\top\right\|_F \\
&\le \left\|FF^\top - I\right\|_F + \|F_{Q\setminus J}\|_F^2 \\
&\le 2|Q|^2\varepsilon_{\mathrm{mom}} + \|F_{Q\setminus J}\|_F^2. && \text{(by (146))} \quad (156)
\end{aligned}$$

We now bound the error from replacing F_J by Z. Since $\|F_J\|_{\mathrm{op}} \le \|F\|_{\mathrm{op}} \le 2$ by (147),

$$\begin{aligned}
&\left\| F_J \operatorname{diag}_{i\in J}(\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle) F_J^\top - Z \operatorname{diag}_{i\in J}(\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle) Z^\top \right\|_F \\
&\quad = \Big\| (F_J - Z) \operatorname{diag}_{i\in J}(\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle) F_J^\top \\
&\qquad + Z \operatorname{diag}_{i\in J}(\langle \boldsymbol{b}, \boldsymbol{a}_i\rangle)(F_J - Z)^\top \Big\|_F \\
&\quad \le K^3(\|F_J\|_{\mathrm{op}} + 1)\, \|F_J - Z\|_F\, \|\boldsymbol{b}\|_2 \qquad (Z \text{ orthogonal; } \|\boldsymbol{a}_i\|_2 \le K^3) \quad (157) \\
&\quad \le 3K^3 \left(2|Q|^2 \varepsilon_{\mathrm{mom}} + \|F_{Q\setminus J}\|_F^2\right) \|\boldsymbol{b}\|_2 . \qquad (\text{by (147) and (156)}) \quad (158)
\end{aligned}$$

This bounds the third term with the same basis for every $\boldsymbol{b}$. □

Proposition D.5 (Spherical Anti-Concentration). *Let $\boldsymbol{b}$ be uniform on the unit sphere $\mathbb{S}^{n-1}$. For every nonzero $\boldsymbol{v} \in \mathbb{R}^n$ and $0 < t < 1$,*

$$\mathbb{P}_{\boldsymbol{b}}\left(|\langle \boldsymbol{b}, \boldsymbol{v}\rangle| \le \frac{t\,\|\boldsymbol{v}\|_2}{\sqrt{n}} \right) \le t.$$

This is the upper bound in [DKM09, Lemma 4], after scaling $\boldsymbol{v}$.

We use the Davis–Kahan sine theorem in the following form.

Proposition D.6 (Davis–Kahan Eigenvector Bound; [YWS15, Corollary 3]). *Let $H, \widetilde{H} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues in nonincreasing order. Fix $j \in [p]$ and $g > 0$ such that*

$$|\lambda_j(H) - \lambda_h(H)| \ge g \qquad (h \in [p] \setminus \{j\}).$$

Let $\boldsymbol{x}, \widetilde{\boldsymbol{x}}$ be unit eigenvectors corresponding to $\lambda_j(H)$ and $\lambda_j(\widetilde{H})$, respectively. The sign of $\widetilde{\boldsymbol{x}}$ can be chosen so that

$$\|\widetilde{\boldsymbol{x}} - \boldsymbol{x}\|_2 \le \frac{2\sqrt{2}\,\left\|\widetilde{H} - H\right\|_{\mathrm{op}}}{g}.$$

D.3 Proof of Theorem 6.1

We verify that the moment, projection, and near-orthogonality bounds give the direction accuracy required by Theorem 6.1.

Proposition D.7 (Observable-Operator Accuracy). *Use the hyperparameters from Hyperparameters 4.1, with $2 \le k \le n$. Fix $0 \le \varepsilon_{\mathrm{mean}} \le e^{-K^2 s}$, and put*

$$\varepsilon_{\mathrm{fam}} = 16K^3k^3 \left((1 + 2K^{s-1})\varepsilon_{\mathrm{mean}} + k e^{-s/K} \right). \quad (159)$$

Then

$$32K^6\sqrt{n}\,k^2\varepsilon_{\mathrm{fam}} \le e^{-s/(2K)} \le \frac{\Delta^2}{16B^2}. \quad (160)$$

Proof. By the choice of Hyperparameters 4.1,

$$K \ge 128, \quad (161)$$
$$s \ge K^4 \log n. \quad (162)$$

The assumed bound on $\varepsilon_{\text{mean}}$ gives

$$\begin{aligned}(1+2K^{s-1})\varepsilon_{\text{mean}} &\le 3e^{-(K^2-\log K)s} && (\varepsilon_{\text{mean}} \le e^{-K^2 s};\ K \ge 1)\\ &\le e^{-s/K}. && (K^2-\log K-1 \ge 1/K,\ 3 \le e^s) \quad (163)\end{aligned}$$

Consequently,

$$\begin{aligned}\varepsilon_{\text{fam}} &\le 16K^3k^3(1+k)e^{-s/K} && (\text{by (159) and (163)})\\ &\le 32K^3k^4e^{-s/K}. && (k \ge 1) \quad (164)\end{aligned}$$

The remaining factor is absorbed by half of the exponential decay:

$$\begin{aligned}1024K^9\sqrt{n}\,k^6 &\le 2^{10+9K}n^7 && (K \le 2^K;\ k \le n)\\ &\le n^{17+9K} && (n \ge 2)\\ &\le n^{K^3/2}. && (\text{choice of } K \text{ (161)}\colon 17+9K \le K^3/2) \quad (165)\end{aligned}$$

Combining these bounds gives

$$\begin{aligned}32K^6\sqrt{n}\,k^2\varepsilon_{\text{fam}} &\le 1024K^9\sqrt{n}\,k^6e^{-s/K} && (\text{by (164)})\\ &\le e^{(K^3/2)\log n - s/K} && (\text{by (165)})\\ &\le e^{-s/(2K)}. && (\text{choice of } s \text{ (162)})\end{aligned}$$

It remains to compare this error bound with the tolerance needed for refinement. The choices in Hyperparameters 4.1 give $s \ge K^4 \log 2$ and $K\Delta^2 \ge 128B^2$, and hence

$$e^{-s/(2K)} \le 2^{-K^3/2} \le K^{-1} \le \frac{\Delta^2}{128B^2} \le \frac{\Delta^2}{16B^2}. \quad (166)$$

Here $2^{-K^3/2} \le K^{-1}$ follows from $K \le 2^K \le 2^{K^3/2}$ for $K \ge 128$. This proves (160). □

E Deferred Proofs for Section 7

Proposition E.1 (Refinement Bounds). *Use K from Hyperparameters 4.1, with $2 \le k \le n$. Fix $\varepsilon \in (0,1)$ and $\delta \in (0, 1/3)$, and set*

$$\Gamma = 2K\sqrt{\log\left(\frac{n}{\varepsilon}\right)}.$$

Then

$$5B\Gamma\sqrt{k}\,e^{-\Gamma^2/K} \le \frac{\varepsilon}{2},$$

$$\frac{n^2B^2\log k}{\varepsilon^2}\Gamma^2 e^{\Gamma^2}\log\left(\frac{n}{\delta}\right) = \widetilde{O}\left(\left(\frac{nk}{\varepsilon}\right)^{5K^2}\log(1/\delta)\right).$$

Proof. We first verify the bias bound, then bound the expression governing the sample complexity and running time. By the choice of Hyperparameters 4.1,

$$K \ge 128, \quad (167)$$

$$B \le K. \quad (168)$$

The choice of Γ gives

$$\Gamma^2 = 4K^2 \log\left(\frac{n}{\varepsilon}\right). \tag{169}$$

Substituting the choice of Γ gives

$$\begin{aligned}
5B\Gamma\sqrt{k}\, e^{-\Gamma^2/K} &= 10BK\sqrt{k\log\left(\frac{n}{\varepsilon}\right)}\left(\frac{n}{\varepsilon}\right)^{-4K} && \text{(choice of } \Gamma \text{ (169))} \\
&= \frac{10BK\sqrt{k}}{n}\sqrt{\log\left(\frac{n}{\varepsilon}\right)}\left(\frac{n}{\varepsilon}\right)^{1-4K}\varepsilon \\
&\le 10K^2\left(\frac{n}{\varepsilon}\right)^{3/2-4K}\varepsilon && (B \le K; k \le n; \log(n/\varepsilon) \le n/\varepsilon) \\
&\le 10K^2\, 2^{3/2-4K}\varepsilon && (n/\varepsilon \ge 2; 3/2 - 4K < 0) \\
&\le 2^{5-2K}\varepsilon && (K \le 2^K; 10 \le 2^{7/2}) \\
&\le \frac{\varepsilon}{2}. && \text{(choice of } K \text{ (167))}
\end{aligned}$$

We now bound the expression governing the sample complexity and running time. The choice of Γ and the ranges of $n, k, \varepsilon, \delta$ give

$$e^{\Gamma^2} = \left(\frac{n}{\varepsilon}\right)^{4K^2} \le \left(\frac{nk}{\varepsilon}\right)^{4K^2}, \tag{170}$$

$$\Gamma^2 \le 4K^2 \log\left(\frac{nk}{\varepsilon}\right), \tag{171}$$

$$\log\left(\frac{n}{\delta}\right) \le 2\log\left(\frac{nk}{\varepsilon}\right)\log(1/\delta). \qquad (\log(1/\delta) \ge 1; \log(k/\varepsilon) \ge \log 2) \tag{172}$$

Since $K \le 2^K$ and $nk/\varepsilon \ge 2$,

$$K^4 \le 2^{4K} \le \left(\frac{nk}{\varepsilon}\right)^{4K}.$$

Using these bounds, $B \le K$, $\log k \le \log(nk/\varepsilon)$, and $n^2/\varepsilon^2 \le (nk/\varepsilon)^2$, we obtain

$$\begin{aligned}
\frac{n^2B^2\log k}{\varepsilon^2}\Gamma^2 e^{\Gamma^2}\log\left(\frac{n}{\delta}\right) &\le 8K^4\left(\frac{nk}{\varepsilon}\right)^{4K^2+2}\log^3\left(\frac{nk}{\varepsilon}\right)\log(1/\delta) \\
&\le 8\left(\frac{nk}{\varepsilon}\right)^{4K^2+4K+2}\log^3\left(\frac{nk}{\varepsilon}\right)\log(1/\delta) && (K^4 \le (nk/\varepsilon)^{4K}) \\
&= \widetilde{O}\left(\left(\frac{nk}{\varepsilon}\right)^{5K^2}\log(1/\delta)\right). && (4K^2 + 4K + 2 \le 5K^2)
\end{aligned}$$

This proves both bounds. □

F Deferred Proofs for Section 8

We give the projected-feature second-moment calculation and the sampling proof of Lemma 8.6. We then prove the joint Gaussian tail bound and the pairwise exceedance bound used for the peeling bias.

F.1 Second-Moment Bounds

Proposition F.1 (Wick's Identity [Iss18]). *Let $\boldsymbol{G} = (G_1, \dots, G_n)$ be a centered jointly Gaussian vector, where $n \geq 1$. If n is odd, then $\mathbb{E}_{\boldsymbol{G}}[\prod_{a=1}^{n} G_a] = 0$. If $n = 2q$, then*

$$\mathbb{E}_{\boldsymbol{G}}\left[\prod_{a=1}^{2q} G_a\right] = \sum_{P \in \mathcal{P}_2(2q)} \prod_{\{a,b\} \in P} \mathbb{E}_{\boldsymbol{G}}[G_a G_b],$$

where $\mathcal{P}_2(2q)$ is the set of partitions of $[2q]$ into unordered pairs.

By multilinearity, the same formula holds for complex linear forms in a real Gaussian vector. Conjugated forms are treated as separate factors in the product.

Proof of Proposition 8.5. Write $\widetilde{\boldsymbol{X}} = \boldsymbol{X} + \mathrm{i}\boldsymbol{X}'$. Let $\mathsf{A} \in (\mathbb{R}^n)^{\otimes 2r}$ be the tensor satisfying $\mathrm{Mat}_r(\mathsf{A}) = U^\top H U$, and put $\mathsf{A}_{\boldsymbol{b}} = \mathsf{A} \otimes \boldsymbol{b}$. Expanding the matrix contractions gives

$$\psi(\boldsymbol{X}, \boldsymbol{X}') = \frac{\operatorname{Re}\left\langle \mathsf{A}, \widetilde{\boldsymbol{X}}^{\otimes 2r} \right\rangle}{\sqrt{(2r)!}}, \qquad \psi_{\boldsymbol{b}}(\boldsymbol{X}, \boldsymbol{X}') = \frac{\operatorname{Re}\left\langle \mathsf{A}_{\boldsymbol{b}}, \widetilde{\boldsymbol{X}}^{\otimes 2r+1} \right\rangle}{\sqrt{(2r+1)!}}.$$

Here the tensor pairing is extended bilinearly to complex tensors. Since $UU^\top = I_p$ and flattening preserves the Frobenius norm,

$$\|\mathsf{A}\|_F = \left\| U^\top H U \right\|_F = \|H\|_F \leq 1, \qquad \|\mathsf{A}_{\boldsymbol{b}}\|_F = \|\mathsf{A}\|_F \, \|\boldsymbol{b}\|_2 \leq \|\boldsymbol{b}\|_2 .$$

The polynomials $\psi(\boldsymbol{X}, \boldsymbol{X}')$ and $\psi_{\boldsymbol{b}}(\boldsymbol{X}, \boldsymbol{X}')$ have degrees at most $2r$ and $2r+1$, respectively.

We first bound the averages over $\boldsymbol{X}'$. The definition of the Hermite tensors gives

$$\mathbb{E}_{\boldsymbol{X}'}[\psi(\boldsymbol{X}, \boldsymbol{X}') \mid \boldsymbol{X}] = \langle \mathsf{H}_{2r}(\boldsymbol{X}), \mathrm{Sym}(\mathsf{A}) \rangle , \qquad \mathbb{E}_{\boldsymbol{X}'}[\psi_{\boldsymbol{b}}(\boldsymbol{X}, \boldsymbol{X}') \mid \boldsymbol{X}] = \langle \mathsf{H}_{2r+1}(\boldsymbol{X}), \mathrm{Sym}(\mathsf{A}_{\boldsymbol{b}}) \rangle .$$

Hermite orthogonality (Proposition 4.5) therefore gives

$$\left\| \mathbb{E}_{\boldsymbol{X}'}[\psi(\boldsymbol{X}, \boldsymbol{X}') \mid \boldsymbol{X}] \right\|_{L^2(\boldsymbol{X})} = \|\mathrm{Sym}(\mathsf{A})\|_F \leq \|\mathsf{A}\|_F \leq 1,$$
$$\left\| \mathbb{E}_{\boldsymbol{X}'}[\psi_{\boldsymbol{b}}(\boldsymbol{X}, \boldsymbol{X}') \mid \boldsymbol{X}] \right\|_{L^2(\boldsymbol{X})} = \|\mathrm{Sym}(\mathsf{A}_{\boldsymbol{b}})\|_F \leq \|\mathsf{A}_{\boldsymbol{b}}\|_F \leq \|\boldsymbol{b}\|_2 .$$

Here averaged symmetrization is the orthogonal projection onto symmetric tensors.

We now bound the sampled functions using Wick's identity (Proposition F.1). The complex Gaussian vector $\widetilde{\boldsymbol{X}}$ satisfies

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'}[\widetilde{X}_j \widetilde{X}_h] = 0, \qquad \mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'}[\widetilde{X}_j \overline{\widetilde{X}_h}] = 2\mathbf{1}_{\{j=h\}}.$$

For a real order-d tensor G, every nonzero Wick pairing in $\mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'} \left| \left\langle \mathsf{G}, \widetilde{\boldsymbol{X}}^{\otimes d} \right\rangle \right|^2$ therefore matches one of the d unbarred factors with one of the d barred factors. Summing over the $d!$ such pairings gives

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'} \left| \left\langle \mathsf{G}, \widetilde{\boldsymbol{X}}^{\otimes d} \right\rangle \right|^2 = 2^d d! \, \langle \mathsf{G}, \mathrm{Sym}(\mathsf{G}) \rangle = 2^d d! \, \|\mathrm{Sym}(\mathsf{G})\|_F^2 \leq 2^d d! \, \|\mathsf{G}\|_F^2 .$$

Since $(\operatorname{Re} z)^2 \leq |z|^2$, applying this identity to A and $\mathsf{A}_{\boldsymbol{b}}$ and dividing by the respective factorials yields

$$\mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'}[\psi(\boldsymbol{X}, \boldsymbol{X}')^2] \leq 2^{2r} \leq 100^{r+2}, \qquad \mathbb{E}_{\boldsymbol{X},\boldsymbol{X}'}[\psi_{\boldsymbol{b}}(\boldsymbol{X}, \boldsymbol{X}')^2] \leq 2^{2r+1} \, \|\boldsymbol{b}\|_2^2 \leq 100^{r+2} \, \|\boldsymbol{b}\|_2^2 .$$

□

F.2 Sampling Error

Proof of Lemma 8.6. Apply Lemma 4.7 to all mp^2 weighted entries, with second-moment bound 100^{s+2}, coordinate accuracy $\varepsilon_{\text{est}}/(p\sqrt{m})$, and failure probability δ. Here $\varepsilon_{\text{est}}/(p\sqrt{m}) \le 1 \le \sqrt{100^{s+2}}$. Thus, with probability at least $1-\delta$, every estimated entry has error at most $\varepsilon_{\text{est}}/(p\sqrt{m})$. Condition on this event and write

$$E_\ell := \widehat{M}_\ell - \mathbb{E}_{\boldsymbol{X},Y_{\max},\boldsymbol{X}'}\left[\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X},Y_{\max})\Psi_\ell(\boldsymbol{X},\boldsymbol{X}')\right].$$

The matrices are symmetric because corresponding upper- and lower-triangular entries have identical observation lists. Consequently,

$$\|E_\ell\|_F = \left(\sum_{a,c=1}^{p} |(E_\ell)_{ac}|^2\right)^{1/2} \le p\,\frac{\varepsilon_{\text{est}}}{p\sqrt{m}} = \frac{\varepsilon_{\text{est}}}{\sqrt{m}}.$$

For every $\boldsymbol{b} \in \mathbb{R}^m$, the triangle inequality gives

$$\begin{aligned}\left\|\sum_{\ell=1}^{m} b_\ell E_\ell\right\|_F &\le \sum_{\ell=1}^{m} |b_\ell|\,\|E_\ell\|_F \\ &\le \|\boldsymbol{b}\|_2 \left(\sum_{\ell=1}^{m} \|E_\ell\|_F^2\right)^{1/2} && \text{(Cauchy–Schwarz)} \\ &\le \varepsilon_{\text{est}}\,\|\boldsymbol{b}\|_2 .\end{aligned}$$

This calculation holds for every $\boldsymbol{b}$ on the same entrywise success event, which proves the simultaneous guarantee.

Since $\log(mp^2/\delta) \le 2\log(mp/\delta)$, the number of samples is

$$O\left(\frac{p^2 m}{\varepsilon_{\text{est}}^2} 100^s \log\left(\frac{mp}{\delta}\right)\right).$$

On each observation, the signed peeling weight can be computed once and reused across the family, and thus each input function is evaluated once. Computing the coordinatewise trimmed means takes time polynomial in the sample count, n, m, p, and $|\widehat{\mathcal{R}}|$, apart from the input-function evaluations. This proves the running-time claim. □

Because $\left|\widehat{\omega}_{\widehat{\mathcal{R}}}(\boldsymbol{X},Y_{\max})\right| \le 1$ and $1 \le r \le s$, Proposition 8.5 verifies the second-moment hypothesis of Lemma 8.6 for every function used by Algorithm 6.1. All entries of a matrix or family use one sample block. Each call uses fresh samples independent of its recovered list and input functions. Theorem 8.1 combines the sampling guarantee with the recovered-regressor and peeling-bias bounds. The call to EstimateProjectedMoments in ImplicitTensorDecomposition uses sampling accuracy $\varepsilon_{\text{est}} = e^{-K^6 s}$.

F.3 Peeling Bias

Proposition F.2 (Joint Gaussian Tail Bound). *Fix a centered jointly Gaussian pair (G_1, G_2) with positive variances σ_1^2, σ_2^2 and covariance c satisfying*

$$D := \sigma_1^2\sigma_2^2 - c^2 > 0, \qquad c \le \min\{\sigma_1^2, \sigma_2^2\}.$$

For every $t > 0$,

$$\mathbb{P}(G_1 > t, G_2 > t) \le \exp\left(-\frac{t^2(\sigma_1^2 + \sigma_2^2 - 2c)}{2D}\right).$$

Proof of Proposition F.2. Fix $T_0 > 0$ and put

$$H := (\sigma_2^2 - c)G_1 + (\sigma_1^2 - c)G_2.$$

The coefficients are nonnegative by the hypothesis on c. They cannot both vanish, since that would give $\sigma_1^2 = \sigma_2^2 = c$ and $D = 0$. Consequently, $\sigma_1^2 + \sigma_2^2 - 2c > 0$, and $G_1 > T_0, G_2 > T_0$ implies $H > T_0(\sigma_1^2 + \sigma_2^2 - 2c)$. The covariance and variance calculations give

$$\begin{aligned}
\operatorname{Cov}(G_1, H) &= (\sigma_2^2 - c)\sigma_1^2 + (\sigma_1^2 - c)c = D, \\
\operatorname{Cov}(G_2, H) &= (\sigma_2^2 - c)c + (\sigma_1^2 - c)\sigma_2^2 = D, \\
\operatorname{Var}(H) &= (\sigma_2^2 - c)\operatorname{Cov}(G_1, H) + (\sigma_1^2 - c)\operatorname{Cov}(G_2, H) \\
&= D(\sigma_1^2 + \sigma_2^2 - 2c) > 0.
\end{aligned}$$

The Gaussian tail bound (Proposition 4.2) gives

$$\begin{aligned}
\mathbb{P}(G_1 > T_0, G_2 > T_0) &\le \mathbb{P}\big(H > T_0(\sigma_1^2 + \sigma_2^2 - 2c)\big) \\
&\le \exp\left(-\frac{T_0^2(\sigma_1^2 + \sigma_2^2 - 2c)^2}{2\operatorname{Var}(H)}\right) && \text{(Gaussian tail bound)} \\
&= \exp\left(-\frac{T_0^2(\sigma_1^2 + \sigma_2^2 - 2c)}{2D}\right).
\end{aligned}$$

This proves the claim. □

Proof of Lemma 8.8. By the choice of Hyperparameters 4.1,

$$K \ge 128, \tag{173}$$

$$K \ge 16(1 + B^2)^3. \tag{174}$$

Fix distinct $i, j \in [k]$ and $1 \le r \le s$. Hölder's inequality separates the polynomial from the probability that both responses exceed T. We use exponent K. Hypercontractivity bounds the polynomial factor, while uncoveredness preserves additional Gaussian decay in the probability factor. Thus

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{X},\xi_i,\xi_j,\boldsymbol{X}'}\big[|\psi(\boldsymbol{X}, \boldsymbol{X}')|\mathbf{1}\{Y_i > T, Y_j > T\}\big] \\
&\quad \le \|\psi\|_{L^K(\boldsymbol{X},\boldsymbol{X}')}\, \mathbb{P}_{\boldsymbol{X},\xi_i,\xi_j}(Y_i > T, Y_j > T)^{1-1/K}. && \text{(Hölder's inequality)} \quad (175)
\end{aligned}$$

We first bound the polynomial factor. By Gaussian hypercontractivity (Proposition 4.4) of the standard Gaussian vector $(\boldsymbol{X}, \boldsymbol{X}')$,

$$\begin{aligned}
\|\psi\|_{L^K(\boldsymbol{X},\boldsymbol{X}')} &\le (K-1)^{(2s+1)/2}\, \|\psi\|_{L^2(\boldsymbol{X},\boldsymbol{X}')} && \text{(by Proposition 4.4)} \\
&\le K^{s+1/2}\, \|\psi\|_{L^2(\boldsymbol{X},\boldsymbol{X}')}. && (176)
\end{aligned}$$

We now bound the probability factor. If Y_i and Y_j were independent, the probability that both responses exceed T would be the product of their Gaussian tail probabilities. For the correlated responses here, uncoveredness supplies the additional factor $e^{-T^2/K}$ needed to absorb the polynomial factors. The pair (Y_i, Y_j) is centered jointly Gaussian, with variances $\sigma_i^2 = 1 + \|\boldsymbol{w}_i\|_2^2$ and $\sigma_j^2 = 1 + \|\boldsymbol{w}_j\|_2^2$, and covariance $c_{ij} = \langle \boldsymbol{w}_i, \boldsymbol{w}_j \rangle$. Cauchy–Schwarz and Assumption 1.2 give

$$D_{ij} := \sigma_i^2 \sigma_j^2 - c_{ij}^2 \geq 1, \qquad \sigma_i^2 - c_{ij},\ \sigma_j^2 - c_{ij} \geq 1 + \Delta^2.$$

Applying the joint Gaussian tail bound (Proposition F.2) and raising the bound to the positive power $1 - 1/K$ gives

$$\begin{aligned} &\mathbb{P}_{\boldsymbol{X}, \xi_i, \xi_j}(Y_i > T, Y_j > T)^{1-1/K} \\ &\quad \leq \exp\left(-T^2 \left(1 - \frac{1}{K}\right) \frac{\sigma_i^2 + \sigma_j^2 - 2c_{ij}}{2D_{ij}} \right). \end{aligned} \tag{177}$$

We next bound the coefficient of $-T^2$.

Using $\sigma_j^2 \leq 1 + B^2$ and $D_{ij} \leq (1 + B^2)^2$,

$$\begin{aligned} &\left(1 - \frac{1}{K}\right) \frac{\sigma_i^2 + \sigma_j^2 - 2c_{ij}}{2D_{ij}} \\ &= \left(1 - \frac{1}{K}\right) \left(\frac{1}{2\sigma_j^2} + \frac{(\sigma_j^2 - c_{ij})^2}{2\sigma_j^2 D_{ij}} \right) \\ &\geq \left(1 - \frac{1}{K}\right) \left(\frac{1}{2\sigma_j^2} + \frac{(1 + \Delta^2)^2}{2(1 + B^2)^3} \right) && \text{(by Assumption 1.2)} \\ &\geq \left(1 - \frac{1}{K}\right) \left(\frac{1}{2\sigma_j^2} + \frac{2}{K} \right) && \text{(choice of } K \text{ (174))} \\ &= \frac{1}{2\sigma_j^2} + \frac{2}{K} - \frac{1}{2K\sigma_j^2} - \frac{2}{K^2} \\ &\geq \frac{1}{2\sigma_j^2} + \frac{1}{K}. && (\sigma_j^2 \geq 1, K \geq 4) \end{aligned} \tag{178}$$

It remains to compare $e^{-T^2/(2\sigma_j^2)}$ with $\lambda_j^{(2r)}$. We use the Gaussian density on the initial interval $[T, T + 1/T]$ of the slice and then apply Lemma 5.8. Since $T^2 \geq Ks \geq 1$ and $\sigma_j^2 \geq 1$,

$$\begin{aligned} \mathbb{P}(T \leq Y_j \leq T + 1) &\geq \int_T^{T+1/T} \phi_{\sigma_j^2}(u)\, \mathrm{d}u \\ &\geq \frac{1}{T} \phi_{\sigma_j^2}(T + 1/T) && \text{(the density decreases)} \\ &= \frac{1}{T} \phi_{\sigma_j^2}(T) \exp\left(-\frac{2 + T^{-2}}{2\sigma_j^2} \right) \\ &\geq \frac{e^{-3/2}}{T} \phi_{\sigma_j^2}(T). && (T \geq 1, \sigma_j^2 \geq 1) \end{aligned} \tag{179}$$

Consequently,

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{X},\xi_i,\xi_j}(Y_i > T, Y_j > T)^{1-1/K} \\
&\quad \le e^{-T^2/K} e^{-T^2/(2\sigma_j^2)} && \text{(by (177) and (178))} \\
&\quad \le e^{-T^2/K} e^{3/2} T\sqrt{2\pi\sigma_j^2}\,\mathbb{P}(T \le Y_j \le T+1) && \text{(by (179))} \\
&\quad \le K^{3/2} T e^{-T^2/K}\mathbb{P}(T \le Y_j \le T+1) && (\sqrt{2\pi\sigma_j^2} \le K,\ e^{3/2} \le \sqrt{K}) \\
&\quad \le K^2 e^{-T^2/(2K)}\mathbb{P}(T \le Y_j \le T+1) && (T/\sqrt{K} \le e^{T^2/(2K)}) \\
&\quad \le K^2 e^{-T^2/(2K)} \lambda_j^{(2r)}. && \text{(by Lemma 5.8)} \quad (180)
\end{aligned}$$

Substituting the two factor bounds into (175) gives

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X},\xi_i,\xi_j,\boldsymbol{X}'}\left[|\psi(\boldsymbol{X},\boldsymbol{X}')|\mathbf{1}\{Y_i > T, Y_j > T\}\right] &\le K^{s+5/2} e^{-T^2/(2K)}\,\|\psi\|_{L^2(\boldsymbol{X},\boldsymbol{X}')}\,\lambda_j^{(2r)} \\
&\le e^{4s\log K - T^2/(2K)}\,\|\psi\|_{L^2(\boldsymbol{X},\boldsymbol{X}')}\,\lambda_j^{(2r)}. \qquad (s \ge 1)
\end{aligned}$$

This proves Lemma 8.8. □

G Deferred Proofs for Section 9

Proposition G.1 (Complexity of Full Recovery). *Fix $\varepsilon \in (0,1)$ and $\delta \in (0, 1/3)$, and use the hyperparameters from Hyperparameters 4.1. Consider at most k calls to each of* ImplicitTensorDecomposition *and* RestrictedLeastSquares, *satisfying the input hypotheses of Theorems 6.1 and 7.1. Use failure probability $\delta/(2k)$ in each call and accuracy ε_{int} for refinement. Their total sample complexity is*

$$\widetilde{O}\left(\left(\frac{nk}{\varepsilon}\right)^{K^{13}} \log(1/\delta)\right),$$

and their total running time is

$$\left(\frac{nk}{\varepsilon}\right)^{\text{poly}(K)} \log(1/\delta).$$

Proof. We substitute the requested accuracy into the bounds for each routine, then sum over the calls. By the choice of Hyperparameters 4.1,

$$K \ge 128, \tag{181}$$

$$s \le 2K^4 \log n, \tag{182}$$

$$\varepsilon_{\text{int}} = \min\{\varepsilon, e^{-K^6 s}\}. \tag{183}$$

The sample calculation for Theorem 6.3 gives the explicit bound (64). The observable-operator routines use no further observations, so this also bounds the sample count of each decomposition call. Replacing the failure probability by $\delta/(2k)$ adds only a logarithmic factor in k, which is suppressed by $\widetilde{O}$.

To bound the samples used by refinement, we first compare the requested accuracy with the final accuracy:

$$\begin{aligned}
\log\left(\frac{nk}{\varepsilon_{\mathrm{int}}}\right) &= \log(nk) + \max\left\{\log\left(\frac{1}{\varepsilon}\right), K^6 s\right\} && \text{(choice of } \varepsilon_{\mathrm{int}} \text{ (183))} \\
&\le (1 + 2K^{10}) \log\left(\frac{nk}{\varepsilon}\right) && \text{(choice of } s \text{ (182))} \\
&\le 3K^{10} \log\left(\frac{nk}{\varepsilon}\right). && (184)
\end{aligned}$$

Thus Theorem 7.1 bounds the samples used by each refinement call by

$$\widetilde{O}\left(\left(\frac{nk}{\varepsilon}\right)^{15K^{12}} \log(1/\delta)\right). \tag{185}$$

The exponent is $15K^{12} = 5K^2 \cdot 3K^{10}$ by (184).

It remains to sum the sample bounds over at most k calls of each kind. Using (64) and (185), the total sample count is

$$\widetilde{O}\left(k\left[(nk)^{4K^{10}+12K^4} + \left(\frac{nk}{\varepsilon}\right)^{15K^{12}}\right]\log(1/\delta)\right).$$

Since $k \le nk/\varepsilon$ and $4K^{10} + 12K^4 \le 15K^{12}$, we have

$$\begin{aligned}
&k\left[(nk)^{4K^{10}+12K^4} + \left(\frac{nk}{\varepsilon}\right)^{15K^{12}}\right] \\
&\quad\le 2\left(\frac{nk}{\varepsilon}\right)^{15K^{12}+1} \\
&\quad\le 2\left(\frac{nk}{\varepsilon}\right)^{K^{13}}. && \text{(choice of } K \text{ (181): } 15K^{12}+1 \le K^{13}) \quad (186)
\end{aligned}$$

This proves the sample bound.

Applying (184) to the running-time bounds in Theorems 6.1 and 7.1 gives

$$\left(\frac{nk}{\varepsilon}\right)^{\mathrm{poly}(K)} \log(1/\delta)$$

operations in total. Here the factor k for the number of calls and the suppressed logarithmic factors are absorbed into the polynomial exponent. □

H AI Methodology

We used `gpt-5.6-sol` extensively for brainstorming, calculations, and literature search as well as `gpt-6-astra`, `claude-fable-5.1` to check for correctness and refine the presentation of the proofs. The author verified the correctness and originality of all content, including references.

More specifically, we were inspired to study conditional moments of regressors $\boldsymbol{X} \mid Y_{\max}$ from prior works on self-selection that studied conditional moments of responses $Y_{\max} \mid \boldsymbol{X}$ [CDIZ23; GM24; KMZ26]. Based on the single-response identity for conditional Hermite moments [Lan57; WN14; Sev26], we developed an explicit formula expressing the restricted moments as approximate sums of nearly orthogonal tensors.

In order to exploit this property and design a polynomial-time algorithm, we turned to prior works that use tensor and spectral methods for parameter estimation in mixture models [LRA93; AHK12; HK13; BCMV14; GVX14; KCL15; LL24; DK25]. Based on these prior works, `gpt-5.6-sol` designed an algorithm for the special case where all regressors (roughly) have the same length. To circumvent this, we suggested recovering the regressors with the longest length first, and `gpt-5.6-sol` was able to generalize this idea to iteratively recover subsets of normalized regressors. Along the way, it designed a refinement procedure that recovers full regressors from accurate estimates of the normalized regressors and a way to simulate the restricted moments of the remaining regressors.