

488 A Limitations

489 In this section we describe some of the primary limitations of our work. First, this paper focuses on
 490 the simple setting of isotropic Gaussian data. Extending both the information leap and generative leap
 491 to more complicated data distributions is left to future work. Next, we focus on deriving estimators
 492 that work with minimal information about the multi-index model P , and which succeed with the
 493 optimal sample complexity in the ambient dimension d . As a result, our sample complexity guarantees
 494 scale with constants $C(P)$ which could potentially be exponentially large in the hidden dimension r .
 495 Finally, we focus primarily on subspace estimation, as it is a requirement for full end-to-end learning.

496 B Proofs of Section 2

497 *Proof of Lemma 1* This is a direct generalization of [Damian et al., 2024a, Lemma D.1]. The k -th
 498 Hermite expansion of the likelihood ratio, viewed as a function of the label \bar{Y}_S , is directly

$$\mathbb{E}_{P_S} \left[\frac{dP}{dP_S}(\bar{Z}_S, \bar{Y}_S) h_k(\bar{Z}_S) | \bar{Y}_S \right] = \mathbb{E}_P[h_S(\bar{Z}_S) | \bar{Y}_S] = \zeta_{k,S}, \quad (6)$$

499 and thus, in $L^2(P_S)$, we have

$$\frac{dP}{dP_S}(\bar{z}_S, \bar{y}_S) = \sum_k \langle \zeta_{k,S}(\bar{y}_S), h_k \bar{z}_S \rangle. \quad (7)$$

501 *Proof of Lemma 2* By orthogonality of Hermite polynomials:

$$\chi^2(P||P_S) = \mathbb{E}_{P_S} \left[\frac{dP}{dP_S}[X, Y]^2 \right] - 1 = \sum_{k \geq 1} \mathbb{E}_{Y, Z_S} [\|\zeta_k(Y; Z_S)\|^2] = \sum_{k \geq 1} \lambda_k^2(S).$$

503 *Proof of Proposition 1* The first statement follows immediately from the definition. To prove 3, we
 504 need to show that $k^* \leq \max_j k(R_j, R_{j+1})$ for any flag \mathcal{F} . Let \bar{S} the subspace associated with the
 505 generative leap k^* , and let j' be the largest index such that $R_{j'} \subseteq \bar{S}$.

506 We claim that for any k and any pair of subspaces $T \subseteq T'$, we have $\text{span}(\Lambda_k(T)) \subseteq T' \cup$
 507 $\text{span}(\Lambda_k(T'))$. Indeed, writing $Y' \in T'$ as $Y' = (Y, \tilde{Y})$ with $Y \in T$ and $\tilde{Y} \in T' \setminus T$, we have
 508 $\zeta_{k,T}(Y) = \mathbb{E}_{\tilde{Y}} \zeta_{k,T'}(Y, \tilde{Y})$ when restricted to $(T')^\perp$ (a subset of T^\perp). Now, suppose towards contra-
 509 diction that $k_{j'} = k(R_{j'}, R_{j'+1}) < k^*$. Since $R_{j'} \subseteq \bar{S}$, we have $\text{span}(\Lambda_k(R_{j'})) \subseteq \bar{S} \cup \text{span}(\Lambda_k(\bar{S}))$
 510 for $k \leq k_{j'}$. But from the definition of k^* we have $\text{span}(\Lambda_k(\bar{S})) = \emptyset$ for $k \leq k_{j'}$, which implies that
 511 $R_{j'+1} \subseteq \bar{S}$, which is a contradiction.

513 *Proof of Proposition 2* The proof is an extension of [Damian et al., 2024a, Prop 2.6]. To prove
 514 $k(S)[P] \leq \inf_{\mathcal{T} \in L^2(P_{\bar{y}_S})} l(S)[(\text{Id}_z \otimes \mathcal{T}_{\bar{y}})_\# P]$, consider $k < k(S)$ and any $\mathcal{T} \in L^2(P_{\bar{y}})$. We have

$$\mathbb{E}[\mathcal{T}(Y, Z_S) \mathbf{h}_k(\bar{Z}_S) | Z_S] = \mathbb{E}_Y[\mathbb{E}[\mathcal{T}(Y, Z_S) \mathbf{h}_k(\bar{Z}_S) | Y, Z_S]] = \mathbb{E}_Y(\mathcal{T}(\bar{Y}_S) \zeta_{k,S}(\bar{Y}_S)) = 0, \quad (8)$$

515 since $\zeta_{k,S} = 0$. To prove $k(S)[P] \geq \inf_{\mathcal{T} \in L^2(P_{\bar{y}})} l(S)[(\text{Id}_z \otimes \mathcal{T}_{\bar{y}})_\# P]$, consider $\mathcal{T} = (\zeta_{k(S),S})_\beta$,
 516 where β is a multiindex such that $(\zeta_{k(S),S})_\beta \neq 0$ (this β must exist by definition of $k(S)$). We verify
 517 that

$$\mathbb{E}(\mathcal{T}(\bar{Y}) H_\beta(\bar{Z})) = \mathbb{E}_{\bar{Y}}[\mathcal{T}(\bar{Y})(\zeta_{k(S),S})_\beta(\bar{Y})] \quad (9)$$

$$= \mathbb{E}_{\bar{Y}}[|(\zeta_{k(S),S})_\beta(\bar{Y})|^2] > 0, \quad (10)$$

518 which shows that $\tilde{\zeta}_{k(S),S} \neq 0$ for the model with label transformation \mathcal{T} .

Proof of Proposition 3 Consider the flag $\tilde{F} = \{\emptyset, \tilde{S}_1, \dots, \tilde{S}_J = \mathbb{R}^r\}$ associated with the leap information exponent. We claim that $k(\tilde{S}_j, \tilde{S}_{j+1}) \leq l(\tilde{S}_j)$ for all $j \in [J]$, or equivalently that

$$\text{span}(\tilde{\Lambda}_k(\tilde{S}_j)) \subseteq \text{span}(\Lambda_k(\tilde{S}_j))$$

for $k \leq l(\tilde{S}_j)$. Indeed, observe that $\tilde{\zeta}_{k,S} = \mathbb{E}_Y[Y \zeta_{k,S}]$. As a consequence, from Proposition 1 we have that $k^* \leq \max_j l(\tilde{S}_j) = l^*$.

521

522 C Proofs of Section 3

523 *Proof of Theorem 1* Let $\mathcal{R}_{\tilde{W}} := \frac{d\mathbb{P}_{\tilde{W}}}{d\mathbb{P}_0}$ denote the likelihood ratio conditioned on \tilde{W} . We begin by
524 computing the full likelihood ratio:

$$\mathcal{R}((x_1, y_1), \dots, (x_n, y_n)) = \frac{\mathbb{E}_{\tilde{W}} [\prod_{i=1}^n \mathbb{P}_{\tilde{W}}[x_i, y_i]]}{\prod_{i=1}^n \mathbb{P}[x_i] \mathbb{P}[y_i]} = \mathbb{E}_{\tilde{W}} \left[\prod_{i=1}^n \mathcal{R}_W(x_i, y_i) \right].$$

525 Then by Lemma 1, we can expand this as

$$\mathcal{R} = \mathbb{E}_{\tilde{W}} \left[\prod_{i=1}^n \left(\sum_{k \geq 0} \langle \zeta_k(\bar{y}_i), \mathbf{h}_k(\tilde{W}^\top \bar{x}_i) \rangle \right) \right].$$

526 We will isolate the low degree part with respect to $\{\bar{x}_1, \dots, \bar{x}_n\}$, which we denote by $\mathcal{R}_{\leq D}$. To
527 compute this, we need to switch the product and the summation:

$$\mathcal{R} = \mathbb{E}_{\tilde{W}} \left[\sum_{p=0}^{\infty} \sum_{k_1 + \dots + k_n = p} \left(\prod_{i=1}^n \langle \zeta_{k_i}(\bar{y}_i), \mathbf{h}_{k_i}(\tilde{W}^\top \bar{x}_i) \rangle \right) \right].$$

528 We note that each term on the right hand side is a polynomial in $\bar{x}_1, \dots, \bar{x}_n$ of degree p which is
529 orthogonal to all polynomials of degree less than p . Therefore $\mathcal{R}_{\leq D}$ is given by:

$$\mathcal{R}_{\leq D} = \mathbb{E}_{\tilde{W}} \left[\sum_{p=0}^D \sum_{k_1 + \dots + k_n = p} \left(\prod_{i=1}^n \langle \zeta_{k_i}(\bar{y}_i), \mathbf{h}_{k_i}(\tilde{W}^\top \bar{x}_i) \rangle \right) \right].$$

530 We can now use the orthogonality property of Hermite polynomials to compute the norms with
531 respect to the null distribution \mathbb{P}_0 . If \tilde{W}, \tilde{W}' are independent draws from the prior on \tilde{W} then:

$$\begin{aligned} \|\mathcal{R}_{\leq D}\|_{L^2(\mathbb{P}_0)}^2 &= \mathbb{E}_{\tilde{W}, \tilde{W}'} \left[\sum_{p=0}^D \sum_{k_1 + \dots + k_n = p} \left(\prod_{i=1}^n \mathbb{E}_{\mathbb{P}_0} \langle \zeta_{k_i}(\bar{y}_i) \otimes \zeta_{k_i}(\bar{y}_i), \mathbf{h}_{k_i}(\tilde{W}^\top \bar{x}_i) \otimes \mathbf{h}_{k_i}((\tilde{W}')^\top \bar{x}_i) \rangle \right) \right] \\ &= \mathbb{E}_{\tilde{W}, \tilde{W}'} \left[\sum_{p=0}^D \sum_{k_1 + \dots + k_n = p} \left(\prod_{i=1}^n \langle \mathbb{E}[\zeta_{k_i} \otimes \zeta_{k_i}], \mathbb{E}[\mathbf{h}_{k_i}(\tilde{W}^\top \bar{x}_i) \otimes \mathbf{h}_{k_i}((\tilde{W}')^\top \bar{x}_i)] \rangle \right) \right]. \end{aligned}$$

532 For a pair $\Sigma, \tilde{\Sigma}$ of operators where Σ is PSD, observe that

$$|\langle \Sigma, \tilde{\Sigma} \rangle| \leq \text{Tr}\{\Sigma\} \|\tilde{\Sigma}\|_{\text{op}},$$

533 thus

$$\|\mathcal{R}_{\leq D}\|_{L^2(\mathbb{P}_0)}^2 \leq \mathbb{E}_{\tilde{W}, \tilde{W}'} \left[\sum_{p=0}^D \sum_{k_1 + \dots + k_n = p} \left(\prod_{i=1}^n \lambda_k^2 \left\| \mathbb{E}[\mathbf{h}_{k_i}(\tilde{W}^\top \bar{x}_i) \otimes \mathbf{h}_{k_i}((\tilde{W}')^\top \bar{x}_i)] \right\| \right) \right].$$

534 Now, let $M = \tilde{W}^\top \tilde{W}' \in \mathbb{R}^{r \times r}$. We have the following control on the Hermite correlation term:

Lemma 5.

$$\left\| \mathbb{E}[\mathbf{h}_k(\tilde{W}^\top \bar{x}) \otimes \mathbf{h}_k((\tilde{W}')^\top \bar{x})] \right\|_{\text{op}} = \|M\|_{\text{op}}^k. \quad (11)$$

535 Let z be a random variable with distribution $\|\tilde{W}^\top \tilde{W}'\|$, where \tilde{W}, \tilde{W}' are drawn independently from
 536 the uniform prior on \tilde{W} , and let $\mathcal{P}_{\leq D}$ be the projection operator onto polynomials of degree at most
 537 D in z . Note that z is subgaussian satisfying $\mathbb{P}\left(b\sqrt{\frac{1}{d}} \leq z \leq a\sqrt{\frac{1}{d}}\right) \leq 1 - c_r b - \bar{c}_r e^{-a^2/4}$ for
 538 explicit constants c_r, \bar{c}_r ; see eg [Bietti et al., 2023, Lemma 3.14]. Then we can upper bound the
 539 above expression as:

$$\|\mathcal{R}_{\leq D}\|_{L^2(\mathbb{P}_0)}^2 \leq \mathbb{E}_z \left[\mathcal{P}_{\leq D} \left[\left(\sum_{k \geq 0} \lambda_k^2 z^k \right)^n \right] \right].$$

540 By linearity of expectation and of the projection operator $\mathcal{P}_{\leq D}$, we can expand this using the binomial
 541 theorem:

$$\|\mathcal{R}_{\leq D}\|_{L^2(\mathbb{P}_0)}^2 \leq \sum_{j \geq 0} \binom{n}{j} \mathbb{E} \left[\mathcal{P}_{\leq D} \left[\left(\sum_{k \geq k^*} \lambda_k^2 z^k \right)^j \right] \right].$$

542 We can further upper bound this expression by using that $\lambda_k^2 \leq \binom{r+k-1}{k}$ (Lemma 14). Plugging this
 543 in for $k \geq k^*$ gives:

$$\begin{aligned} \|\mathcal{R}_{\leq D}\|_{L^2(\mathbb{P}_0)}^2 - 1 &\lesssim \sum_{j=1}^{\lfloor D/k^* \rfloor} \binom{n}{j} \mathbb{E}_z \left[\mathcal{P}_{\leq D} \left[\left(\sum_{k \geq k^*} k^{r-1} z^k \right)^j \right] \right] \\ &\lesssim \sum_{j=1}^{\lfloor D/k^* \rfloor} \binom{n}{j} \mathbb{E}_z \left[\mathcal{P}_{\leq D} \left[(k^*)^{j(r-1)} z^{jk^*} (1-z)^{-jr} \right] \right] \\ &\lesssim \sum_{j=1}^{\lfloor D/k^* \rfloor} \binom{n}{j} \left[(k^*)^{j(r-1)} \mathbb{E}_z [z^{jk^*}] \right], \end{aligned}$$

544 where the last line follows from Lemma 6. Finally, since z is $\Theta(\sqrt{1/d})$ -subgaussian, we have
 545 $\mathbb{E}[z^{jk^*}] \lesssim (jk^*/d)^{jk^*/2}$.

546 Now, if $n = O(d^{k^*/2-\gamma})$ with $\gamma > 0$ and $D = O((\log d)^2)$, we have

$$\begin{aligned} \|\mathcal{R}_{\leq D}\|_{L^2(\mathbb{P}_0)}^2 - 1 &\lesssim \sum_{j=1}^{\lfloor D/k^* \rfloor} \binom{n}{j} \left[(k^*)^{j(r-1)} (jk^*/d)^{jk^*/2} \right] \\ &\lesssim n^{D/k^*} k^{*(r-1)D/k^*} (D/d)^{D/2} \\ &= k^{*(r-1)D/k^*} (D)^{D/2} (n^{1/k^*} d^{-1/2})^D \\ &= o_d(1). \end{aligned}$$

547 ■

548 *Proof of Lemma 5* Let $\mathbf{M} \in \mathbb{R}^{d^k \times d^k}$ be the matrix representation of $\mathbb{E}[\mathbf{h}_k(\tilde{W}^\top \tilde{x}) \otimes \mathbf{h}_k((\tilde{W}')^\top \tilde{x})]$.
 549 Let $\mathcal{H}_k \subset L^2(\mathbb{R}^r, \gamma)$ be the space spanned by harmonics of degree k . Observe that for $f, \tilde{f} \in \mathcal{H}_k$,
 550 $f = \sum_{|\beta|=k} c_\beta h_\beta$, $\tilde{f} = \sum_{|\beta|=k} \tilde{c}_\beta h_\beta$ with $c_\beta = \langle f, h_\beta \rangle$, $\tilde{c}_\beta = \langle \tilde{f}, h_\beta \rangle$ we have

$$c^\top \mathbf{M} \tilde{c} = \langle P_W f, P_{W'} \tilde{f} \rangle_{\gamma_d}, \quad (12)$$

551 where $P_W f(x) = f(W^\top x)$. We deduce that \mathbf{M} is the ‘averaging operator’ \mathbf{A}_M from [Bietti et al.,
 552 2023, Definition 1.1], restricted at harmonic k . From the SVD of $M = U \Lambda V^\top$, we have [Bietti et al.,
 553 2023, Corollary 2.8] that

$$\mathbf{M} = \sum_{|\beta|=k} \lambda^\beta H_\beta(U) \otimes H_\beta(V), \quad (13)$$

554 with $\lambda^\beta = \prod_j \lambda_j^{\beta_j}$. We thus conclude that $\|\mathbf{M}\|_{\text{op}} = \lambda_{\max}^k = \|M\|^k$. ■

555 **Lemma 6.** Let $z = \|W^\top W'\|_{\text{op}}$, where W, W' are drawn iid from the Haar measure of $\mathcal{S}(r, d)$.
 556 Then, for $l, \tilde{l} \leq d/4$, we have

$$\mathbb{E}_z \left[z^l (1 - z)^{-\tilde{l}} \right] \lesssim \mathbb{E}_z [z^l] . \quad (14)$$

557 *Proof.* The proof is adapted from [Damian et al., 2023, Lemma 26] to the $r > 1$ setting. From [Bietti
 558 et al., 2023, Eq (197)], the joint distribution of singular values $0 \leq \lambda_r \leq \lambda_{r-1} \leq \dots \leq \lambda_1$ of M is
 559 given by

$$p_{r,d}(\lambda_1, \dots, \lambda_r) = Z_{r,d}^{-1} \prod_{i < j} (\lambda_i^2 - \lambda_j^2) \prod_{i=1}^r (1 - \lambda_i^2)^{(d-2r-1)/2} \mathbf{1}(0 \leq \lambda_r \leq \dots \leq \lambda_1 \leq 1) , \quad (15)$$

560 with $Z_{r,d} = \frac{\Gamma_r^2(r/2)}{\pi^{r^2/2}} \frac{\Gamma_r((d-r)/2)}{\Gamma_r(d/2)}$. We have

$$\mathbb{E} \left(z^l (1 - z)^{-\tilde{l}} \right) = \int \lambda_1^l (1 - \lambda_1)^{-\tilde{l}} p_{r,d}(\lambda_1, \dots, \lambda_r) d\lambda_1 \dots d\lambda_r \quad (16)$$

561 From $\lambda_1 \leq 1$ we have $1 - \lambda_1^2 \leq 2(1 - \lambda_1)$ so $(1 - \lambda_1)^{-\tilde{l}} \leq 2^{\tilde{l}} (1 - \lambda_1^2)^{-\tilde{l}}$, thus

$$\mathbb{E} \left(z^l (1 - z)^{-\tilde{l}} \right) \leq 2^{\tilde{l}} \int \lambda_1^l (1 - \lambda_1^2)^{-\tilde{l}} p_{r,d}(\lambda_1, \dots, \lambda_r) d\lambda_1 \dots d\lambda_r \quad (17)$$

$$= 2^{\tilde{l}} Z_{r,d}^{-1} \int \lambda_1^l \prod_{i < j} (\lambda_i^2 - \lambda_j^2) \prod_{i=1}^r (1 - \lambda_i^2)^{(d-2r-1-2\tilde{l})/2} \mathbf{1}(0 \leq \lambda_r \leq \dots \leq \lambda_1 \leq 1) d\lambda_1 \dots d\lambda_r \quad (18)$$

$$= 2^{\tilde{l}} \frac{Z_{r,d-2\tilde{l}}}{Z_{r,d}} \mathbb{E}_{\tilde{z}} [z^l] , \quad (19)$$

562 where \tilde{z} is the largest singular value of $M = W^\top W'$ with $W, W' \in \mathcal{S}(r, d - 2\tilde{l})$. For $d \gg 1$, we
 563 thus conclude that $\mathbb{E} \left(z^l (1 - z)^{-\tilde{l}} \right) \lesssim \mathbb{E}_z [z^l]$. ■

564 D Proofs of Section 4

565 **Lemma 3.** If K is integrally strictly positive definite,⁵ there exist $c(\mathbf{P}, K), C(\mathbf{P}, K) > 0$ independent
 566 of d such that if $S := (U^\star)^\top \text{span}[\Lambda_k]$ denotes the subspace corresponding to the next leap then

$$c(\mathbf{P}, K) \Pi_S \preceq \mathbb{E} U_n \preceq C(\mathbf{P}, K) \Pi_S.$$

567 *Proof.* Let \mathcal{K} be the kernel operator:

$$(\mathcal{K}f)(y) = \mathbb{E}_Y [K(Y, y) f(y)].$$

568 Using that $\langle \mathbb{E}[\mathbf{h}_k(X)|Y], v^{\otimes k} \rangle = \langle \zeta_k(Y), u^{\otimes k} \rangle$ where $u = U^\star v \in \mathbb{R}^r$, we have that for any v :

$$\begin{aligned} v^\top \mathbb{E} M_n v &= u^\top \mathbb{E} [\text{Mat}_{(1,k-1)}[\zeta_k(Y)] \text{Mat}_{(1,k-1)}[\zeta_k(Y')]^\top K(Y, Y')] u \\ &= \langle \zeta_k(\cdot)[u], \mathcal{K} \zeta_k(\cdot)[u] \rangle . \end{aligned}$$

569 First, because $K(y, y) \leq 1$ we have that $\|\mathcal{K}\|_{\text{op}} \leq 1$ so this is upper bounded by

$$\mathbb{E}_Y \|\zeta_k(Y)[u]\|^2 \leq \lambda_k^2 \|u\|^2 .$$

570 Therefore $\mathbb{E} M_n \preceq C(\mathbf{P}, K) \Pi_S$ with $C(\mathbf{P}, K) = \lambda_k^2$. Next, let $v \in S$ with $\|v\| = 1$ so that
 571 $\zeta_k(Y)[v] \neq 0 \in L^2(\mathbf{P}_y)$. Then because K is injective we have that

$$c(v) := \langle \zeta_k(\cdot)[v], \mathcal{K} \zeta_k(\cdot)[v] \rangle > 0.$$

572 Therefore by compactness, if $C(\mathbf{P}, K)$ denotes the minimum value of $c(v)$ over the unit vectors in
 573 S , we have that $C(\mathbf{P}, K) > 0$. In addition we have that $\mathbb{E} M_n \succeq C(\mathbf{P}, K) \Pi_S$ which completes the
 574 proof. ■

⁵We say that K is integrally strictly positive definite if for all finite non-zero signed Borel measures μ ,
 $\int K(x, y) d\mu(x) d\mu(y) > 0$. We remark that many commonly used kernels, including the RBF and Laplacian
 kernels, satisfy this assumption [Sriperumbudur et al., 2010].

575 **Theorem 2.** Let K be a PSD kernel with $K(y, y) \leq 1$ for all y . Then if $n \geq d \text{ polylog}(1/\delta)$ we
 576 have with probability at least $1 - \delta$,

$$\|U_n - \mathbb{E} U_n\| \lesssim_k \frac{d^{k/2}}{n} + r^{k/2} \sqrt{\frac{d}{n}}.$$

577 *Proof.* Note that by the standard decoupling argument (de la Peña and Giné [1999], Theorem 3.4.1),
 578 it suffices to control the tails of the decoupled U -statistic:

$$M_n := \frac{1}{n^2} \sum_{i,j} \phi(x_i) \phi(x'_j)^T K(y_i, y'_j)$$

579 where $\{(x'_i, y'_i)\}_{i=1}^n$ are an i.i.d. copy of $\{(x_i, y_i)\}_{i=1}^n$. We will begin by applying Corollary 2 with
 580 respect to the randomness in $\{(x_i, y_i)\}_{i=1}^n$, treating the replicas $\{(x'_i, y'_i)\}_{i=1}^n$ as fixed. Define

$$V'_n(Y) := \frac{1}{n} \sum_{j=1}^n \phi(x'_j) K(Y, y'_j)$$

581 so that

$$M_n = \frac{1}{n} \sum_i \phi(x_i) V'_n(y_i)^T := \frac{1}{n} \sum_i Z_i.$$

582 Note that if $Z \equiv Z_i$,

$$\|Z\|_{op}^2 \leq \|Z\|_F^2 = \sum_{i,j} \langle \phi(X)^T e_i, V'_n(Y)^T e_j \rangle^2.$$

583 Taking $p/2$ norms and using Lemma 9 gives for $p \geq r \log r$:

$$\begin{aligned} \left\| \|Z\|_{op} \right\|_p^2 &= \left\| \|Z\|_{op} \right\|_{p/2}^2 \\ &\leq \sum_{i,j} \left\| \langle \phi(X)^T e_i, V'_n(Y)^T e_j \rangle \right\|_{p/2}^2 \\ &= \sum_{i,j} \left\| \langle \phi(X)^T e_i, V'_n(Y)^T e_j \rangle \right\|_p^2 \\ &\lesssim_k p^k \sum_{i,j} \left\| \|V'_n(Y)^T e_j\|_F \right\|_{2p}^2 \\ &\leq p^k d^2 \left\| \|V'_n(Y)\|_{op} \right\|_{2p}^2. \end{aligned}$$

584 Next we will compute $\sigma_*(Z)$:

$$\sigma_*(Z)^2 = \sup_{\|u\|=\|v\|=1} \mathbb{E} \left[\langle \phi^T u, V'_n(Y)^T v \rangle^2 \right] \lesssim_k \left\| \|V'_n(Y)\|_{op} \right\|_4^2$$

585 by the same argument as above. Therefore applying Corollary 2 gives that for $p \leq d^c$,

$$\left\| \|M_n - \mathbb{E} M_n\|_{op} \right\|_p \lesssim \left\| \|V'_n(Y)\|_{op} \right\|_{2p} \sqrt{\frac{d}{n}}.$$

586 Now let E' denote the expectation with respect to the replicas $\{(x_i, y_i)\}_{i=1}^n$. Then by Corollary 2:

$$(\mathbb{E}' \mathbb{E} \|U - \mathbb{E} U\|_{op}^p)^{1/p} \lesssim \left(\mathbb{E}' \left\| \|V'_n(Y)\|_{op} \right\|_{2p}^p \right)^{1/p} \sqrt{\frac{d}{n}} = \left(\mathbb{E}_Y \mathbb{E}' \|V'_n(Y)\|_{op}^{2p} \right)^{\frac{1}{2p}} \sqrt{\frac{d}{n}}.$$

587 Now we decompose:

$$\|V(Y)\|_{op} \leq \|\mathbb{E}' V(Y)\| + \|V(Y) - \mathbb{E} V(Y)\|.$$

Because $|K(Y, y'_j)| \leq 1$, we can use Lemma 8 and a standard symmetrization argument to show that the second term has p -norms bounded by $O(\sqrt{\max(d, d^{k-1})/n})$ for $p < d^c$. For the first term we have

$$\|\mathbb{E}' V(Y)\|_{op} \leq \|\mathbb{E}' V(Y)\|_F = \|\mathbb{E}_{Y'} \zeta_k(Y') K(Y, Y')\|_F \leq \sqrt{\mathbb{E}_{Y'} [\|\zeta_k(Y')\|_F^2]} \leq r^{k/2}.$$

Combining everything and applying Markov's inequality gives that with probability at least $1 - \text{poly}(n)e^{-d^c}$,

$$\|U_n - \mathbb{E} U_n\|_{op} \lesssim \left[r^{k/2} + \sqrt{\frac{\max(d, d^{k-1})}{n}} \right] \sqrt{\frac{d}{n}} \lesssim \frac{d^{k/2}}{n} + r^{k/2} \sqrt{\frac{d}{n}}.$$

593

Lemma 4. *If the kernel K is L -Lipschitz, then there exists a constant $C(P, K)$ such that the map $S \rightarrow \mathbb{E} U_n^{(S)}$ is $C(P, K)L$ -Lipschitz in operator norm.*

Proof. Let $Z(Y; X) := \text{Mat}_{(1,k)}[\zeta_k(Y; X_{S \cup S'})]$. Then,

$$\|\mathbb{E} U_n^{(S)} - \mathbb{E} U_n^{(S')}\|_F = \|\mathbb{E}_{X,Y} Z(Y; X) Z(Y'; X')^T E(Y, X)\|_F$$

where

$$E(Y, X) := K((Y, X_S), (Y', X'_S)) - K((Y, X_{S'}), (Y', X'_{S'})).$$

Note that

$$E(Y, X) \leq L \sqrt{\|X_S - X_{S'}\|^2 + \|X'_S - X'_{S'}\|^2} \leq 2d(S, S')L \max(\|X\|, \|X'\|).$$

Then by Holder's inequality,

$$\begin{aligned} \|\mathbb{E} U_n^{(S)} - \mathbb{E} U_n^{(S')}\|_F &\leq \|Z(Y; X)\|_F^2 \|E(Y, X)\|_2 \\ &\leq 2(3r)^k \sqrt{r} L d(S, S') \\ &\lesssim_k r^{k+1} L d(S, S'). \end{aligned}$$

600

Theorem 3. *For any multi-index model P , there exists a constant $C(P, K)$ independent of d such that if $n \geq C(P, K) \left[\frac{d^{k^*/2}}{\epsilon} + \frac{d}{\epsilon^2} + d \text{polylog}(\frac{1}{\delta}) \right]$ then the output $S \subset \mathbb{R}^d$ of Algorithm 2 satisfies $d(S, \text{span}[(U^*)^\top]) \leq \epsilon$ with probability at least $1 - \delta$.*

Proof. Recall the leap decomposition $\mathcal{F} = \{\emptyset = S_0^* \subsetneq S_1^* \subsetneq \dots \subsetneq S_L^* = \mathbb{R}^r\}$. We will prove by induction that for any $\epsilon > 0$, there exists a constant $C(P, K)$ such that the output S_i of Algorithm 2 at step i satisfies $d(S_i, (U^*)^T S_i^*)$ with high probability whenever

$$n \geq C(P, K) \left[\frac{d^{k/2}}{\epsilon} + \frac{d}{\epsilon^2} \right].$$

Note that for $i = 1$ the result is implied directly by Corollary 1. Now assume the result for $i > 1$. By Lemma 3

$$\mathbb{E} U_n^{(U^*)^T S_i^*} \succeq c(P, K) \Pi_{(U^*)^T S_{i+1}^*}.$$

In addition we have by Theorem 2

$$\|U_n^{(S_i)} - \mathbb{E} U_n^{(S_i)}\|_{op} \lesssim_k \frac{d^{k/2}}{n} + r^{k/2} \sqrt{\frac{d}{n}}.$$

Finally by Lemma 4

$$\|\mathbb{E} U_n^{S_i} - \mathbb{E} U_n^{(U^*)^T S_i^*}\|_{op} \lesssim_k r^{k+1} \times L \times d(S_i, (U^*)^T S_i^*).$$

611 Putting it all together we have that:

$$\left\| U_n^{(S_i)} - \mathbb{E} U_n^{(U^*)^T S_i^*} \right\| \lesssim_k \frac{d^{k/2}}{n} + r^{k/2} \sqrt{\frac{d}{n}} + r^{k+1} \times L \times d(S_i, (U^*)^T S_i^*).$$

612 In addition, by the induction hypothesis, $d(S_i, (U^*)^T S_i^*) \leq C(P, K) \left[\frac{d^{k/2}}{n} + r^{k/2} \sqrt{\frac{d}{n}} \right]$. Therefore,

$$\left\| U_n^{(S_i)} - \mathbb{E} U_n^{(U^*)^T S_i^*} \right\| \lesssim_k C(P, K) \left[\frac{d^{k/2}}{n} + r^{k/2} \sqrt{\frac{d}{n}} \right]$$

613 and the result again follows from the Davis-Kahan inequality. ■

614 D.1 Auxiliary Lemmas for Concentration

615 We will start with this simple inequality on the Frobenius norm of a Hermite tensor:

616 **Lemma 7.** $\|h_k(X)\|_F \lesssim_k \|X\|^k + d^{k/4}$.

617 *Proof.* We will use the identity:

$$h_k(X) = \frac{1}{\sqrt{k!}} \mathbb{E}_{Z \sim N(0, I_d)} [(X + iZ)^{\otimes k}].$$

618 Therefore,

$$\begin{aligned} \|h_k(X)\|_F^2 &= \frac{1}{k!} \mathbb{E}_{Z, Z'} [(X + iZ) \cdot (X + iZ')]^k \\ &= \frac{1}{k!} \mathbb{E}_{Z, Z'} [(\|X\|^2 - Z \cdot Z' + iX \cdot (Z + Z'))^k] \\ &\leq \frac{3^{k-1}}{k!} [\|X\|^{2k} + \mathbb{E}_{Z, Z'} [|Z \cdot Z'|^k] + \mathbb{E}_{Z, Z'} [|X \cdot (Z + Z')|^k]] \\ &\lesssim_k \|X\|^{2k} + d^{k/2} + \|X\|^k \\ &\lesssim_k \|X\|^{2k} + d^{k/2}. \end{aligned}$$

619 ■

620 We can use this to concentrate sums of $\sum_{i=1}^n c_i \phi(x_i)$ in operator norm:

621 **Lemma 8.** *There exists an absolute constant C_k such that if $n = d^{1+\epsilon}$ with $\epsilon > 0$ and for any*
 622 *constants c_i with $|c_i| \leq 1$ and $p = d^c$ where $c = \min(1, \epsilon/4)$,*

$$\mathbb{E} \left[\left\| \frac{1}{n} \sum_{i=1}^n c_i \phi(x_i) \right\|_{op}^p \right]^{1/p} \leq C_k \sqrt{\frac{\max(d, d^{k-1})}{n}}.$$

623 *Proof.* Note that for $k = 1, 2$ this follows from the standard bounds for a Gaussian covariance matrix
 624 ($k = 2$) and the norm of a Gaussian vector ($k = 1$). Therefore we will assume $k > 2$. We will begin
 625 by computing $\sigma_*(\phi)$:

$$\sigma_*(\phi) = \sup_{\|u\|=\|v\|=1} \mathbb{E}[(u^T \phi v)^2] = \mathbb{E} \langle \text{vec}[h_k(X)], \text{vec}[u \otimes v] \rangle^2 \leq 1.$$

626 Next, $\|\phi\|_{op} \leq \|\phi\|_F \lesssim_k \|X\|^k + d^{k/4}$. Therefore the p -norms of $\|\phi\|_{op}$ are bounded by $d^{k/2}$ for
 627 any $p \leq d$. Plugging this into Lemma 12 gives that for $p \leq d$,

$$\left\| \left\| \frac{1}{n} \sum_{i=1}^n c_i \phi(x_i) \right\|_{op} \right\|_p \lesssim \sqrt{\frac{d^{k-1}}{n}} + \left(\frac{d^{k/2}}{n} \right)^{1/3} \left(\frac{d^{k-1}}{n} \right)^{1/3} p^{2/3} + \frac{d^{k/2} p}{n}.$$

628 Plugging in $p = d^c$ gives that the second and third terms are dominated by the first which completes
 629 the proof. ■

630 We will also use the following simple lemma:

631 **Lemma 9.** *Let (X, Y) follow a Gaussian single index model with hidden dimension r . Then for any*
 632 *k -tensor-valued random variable $F(Y)$,*

$$\|\langle \mathbf{h}_k(X), F(Y) \rangle\|_p \leq (2p-1)^{\frac{k}{2}} \binom{r+kp}{r}^{\frac{1}{2p}} \|F(Y)\|_F \|_{2p}.$$

633 *Proof.* Without loss of generality we can assume p is even. Now $\langle \mathbf{h}_k(X), f(Y) \rangle^p$ is a polynomial of
 634 degree kp in X . Therefore by Lemma 15

$$\mathbb{E} \langle \mathbf{h}_k(X), F(Y) \rangle^p \leq \sqrt{\mathbb{E}_0 \langle \mathbf{h}_k(X), F(Y) \rangle^{2p} \binom{r+kp}{r}} \leq (2p-1)^{\frac{kp}{2}} \sqrt{\binom{r+kp}{r} \mathbb{E}_Y \|F(Y)\|^{2p}}.$$

635 Taking p th roots gives:

$$\|\langle \mathbf{h}_k(X), F(Y) \rangle\|_p \leq (2p-1)^{\frac{k}{2}} \binom{r+kp}{r}^{\frac{1}{2p}} \|F(Y)\|_F \|_{2p}.$$

636

637 **Lemma 10** (Gaussian hypercontractivity). *Let f be a polynomial of degree k and let $X \sim N(0, I_d)$.*
 638 *Then for $p \geq 2$,*

$$\mathbb{E}_X [|f(X)|^p]^{2/p} \leq (p-1)^k \mathbb{E}_X [f(X)^2].$$

639 **Lemma 11.** *Let X, Y be random variables with $\|Y\|_p \leq Bp^{k/2}$ for*

$$p = \min \left(2, \frac{1}{k} \cdot \log \left(\frac{\|X\|_2}{\|X\|_1} \right) \right).$$

640 *Then,*

$$\mathbb{E}[XY] \leq \|X\|_1 \cdot B \cdot (ep)^{k/2}.$$

641 For any mean zero random matrix Y we define:

$$\begin{aligned} \sigma(Y) &:= \max \left(\|\mathbb{E}[Y Y^\top]\|_2, \|\mathbb{E}[Y^\top Y]\|_2 \right)^{1/2} \\ \sigma_*(Y) &:= \sup_{\|u\|=\|v\|=1} \mathbb{E}[(u^\top Y v)^2] \end{aligned}$$

642 For non-centered matrices, we define $\sigma(Y) := \sigma(Y - \mathbb{E} Y)$ and $\sigma_*(Y) := \sigma_*(Y - \mathbb{E} Y)$.

643 We will rely on the following simple corollary of [Brailovskaya and van Handel 2023, Theorem 2.6]:

644 **Lemma 12.** *Let $Y = \sum_{i=1}^n Z_i$ where Z_i are mean zero independent random matrices. Assume that*
 645 *for all i , $\mathbb{P}[\|Z_i\| > R] \leq \delta$. Then there exists an absolute constant C such that for any $t \geq 0$, with*
 646 *probability at least $1 - n\delta - de^{-t}$,*

$$\|Y\| \leq C \left[\sigma(Y) + \sigma_*(Y) t^{1/2} + R^{1/3} \sigma(Y)^{2/3} t^{2/3} + Rt \right].$$

647 *Proof.* Define $\tilde{Z}_i := Z_i \mathbf{1}_{\|Z_i\|_2 \leq R}$ and let $\tilde{Y} := \sum_{i=1}^n \tilde{Z}_i$. Then,

$$\sigma(\tilde{Y}) = n^{1/2} \sigma(\tilde{Z}) \leq n^{1/2} \sigma(Z) = \sigma(Y)$$

648 and similarly for σ_* . In addition, by definition, $\|\tilde{Z}_i\| \leq R$. Therefore, by [Brailovskaya and van
 649 Handel, 2024, Theorem 2.6] and [Bandeira et al., 2023, Lemma 4.10], there exists a constant C such
 650 that for any $t \geq 0$, with probability at least $1 - de^{-t}$,

$$\|\tilde{Y} - \mathbb{E} \tilde{Y}\| \leq C \left[\sigma(Y) + \sigma_*(Y) t^{1/2} + R^{1/3} \sigma(Y)^{2/3} t^{2/3} + Rt \right].$$

651 Next, note that

$$\|\mathbb{E} Y - \mathbb{E} \tilde{Y}\|_{op} = \left\| \sum_i \mathbb{E}[Z_i \mathbf{1}_{\|Z_i\|_2 > R}] \right\| \leq \sum_i \sigma_*(Z_i) \sqrt{\delta} \leq \sigma_*(Y) \sqrt{n\delta}.$$

652 Now if $\delta > 1/n$, then $1 - n\delta < 0$ so the result is trivially true. Otherwise, $\|\mathbb{E} Y - \mathbb{E} \tilde{Y}\|_{op} \leq \sigma_*(Y)$.

653 Finally, as $\tilde{Y} = Y$ on the event that $\max_i \|Z_i\| \leq R$, a union bound completes the proof. ■

654 We will use the following simple lemmas about σ, σ_* :

655 **Lemma 13.** For any random matrix $A \in \mathbb{R}^{d \times s}$, $\sigma(A)^2 \leq \max(d, s)\sigma_*(A)^2$.

656 *Proof.* Without loss of generality we can assume $\mathbb{E}[A] = 0$. Expanding the definition gives:

$$\sigma(A)^2 = \max(\|\mathbb{E}[AA^\top]\|, \|\mathbb{E}[A^\top A]\|).$$

657 First,

$$\|\mathbb{E}[AA^\top]\| = \sup_{\|v\|=1} \mathbb{E}[v^\top AA^\top v] = \sup_{\|v\|=1} \mathbb{E}[\|A^\top v\|^2] = \sup_{\|v\|=1} \sum_{i=1}^s \mathbb{E}[(e_i^\top A^\top v)^2] \leq s\sigma_*(A)^2.$$

658 Performing the same calculation for A^\top in place of A gives that $\|\mathbb{E}[AA^\top]\| \leq d\sigma_*(A)^2$. Combining
659 these inequalities gives the desired result. ■

660 **Corollary 2.** Let $Y = \frac{1}{n} \sum_{i=1}^N Z_i$ where $Z_i \in \mathbb{R}^{d \times d}$ are mean zero independent random matrices.
661 Assume that for some R, k , $\|Z_i\|_{op} \leq Rt^{k/2}$ with probability at least $1 - e^{-t}$ for all $t \geq 0$. Then if
662 $n = d^{1+\epsilon}$ with $\epsilon > 0$ and $c = \min(1, \frac{\epsilon}{k+4})$, then for all $p \leq d^c$,

$$\mathbb{E}[\|Y\|_{op}^p]^{1/p} \leq C \max\left(\sigma_*(Z), \frac{R}{d}\right) \sqrt{\frac{d}{n}}.$$

663 where C is an absolute constant.

664 *Proof.* First by a union bound, we have that $\max_i \|Z_i\|_2 \lesssim Rt^k$ with probability at least $1 - ne^{-t}$.
665 Substituting this and Lemma 13 into Lemma 12 gives that with probability at least $1 - 2ne^{-t}$,

$$\|Y\| \leq C \max\left(\sigma_*(Z), \frac{R}{d}\right) \left[\sqrt{\frac{d+t}{n}} + \left(\frac{dt^{k/2}}{n}\right)^{1/3} \left(\frac{d}{n}\right)^{1/3} t^{2/3} + \frac{dt^{\frac{k}{2}+1}}{n} \right].$$

666 We can factorize this by pulling out the $\sqrt{d/n}$ and using $n \geq d^{1+\epsilon}$:

$$\|Y\| \leq C \max\left(\sigma_*(Z), \frac{R}{d}\right) \sqrt{\frac{d}{n}} \left[1 + \frac{t^{1/2}}{d^{1/2}} + \frac{t^{\frac{k+4}{6}}}{d^{\frac{\epsilon}{6}}} + \frac{t^{\frac{k}{2}+1}}{n^{\frac{\epsilon}{2}}} \right].$$

667 We can convert this to an p -norm bound for $p \geq \log n$

$$\mathbb{E}[\|Y\|^p]^{1/p} \leq C \max\left(\sigma_*(Z), \frac{R}{d}\right) \sqrt{\frac{d}{n}} \left[1 + \frac{p^{1/2}}{d^{1/2}} + \frac{p^{\frac{k+4}{6}}}{d^{\frac{\epsilon}{6}}} + \frac{p^{\frac{k}{2}+1}}{n^{\frac{\epsilon}{2}}} \right].$$

668 Now if $p = d^c$ where $c = \min(1, \frac{\epsilon}{k+4})$ then the error terms are all less than 1 so we are done. ■

669 **Lemma 14.** For any $p \geq 2$, $\|\zeta_k(Y)\|_F^2 \leq (p-1)^k \binom{r+k-1}{k} \leq ((p-1)r)^k$.

670 *Proof.* By Jensen's inequality and Gaussian hypercontractivity we have

$$\begin{aligned} \|\mathbb{E}[He_k(Z)|Y]\|_F^2 &\leq \mathbb{E}[\|He_k(Z)\|_F^p]^{2/p} \\ &\leq (p-1)^k \mathbb{E}\|He_k(Z)\|_F^2 \\ &= (p-1)^k k! \binom{r+k-1}{k}. \end{aligned}$$

671 Dividing by $k!$ to revert to the normalized Hermite polynomials $\{h_k\}$ completes the proof. ■

672 We will now bound the low-degree density ratio between the joint distribution of (X, Y) and the null
673 distribution $\mathbb{P}_0 := \mathbb{P}_X \otimes \mathbb{P}_Y$:

674 **Lemma 15.** Let $\mathbb{P}_0 := \mathbb{P}_X \otimes \mathbb{P}_Y$ be the null distribution and let $\mathcal{P}_{\leq D}$ denote the orthogonal
675 projection onto polynomials in X of degree at most D . Then:

$$\mathcal{P}_{\leq D} \left(\frac{d\mathbb{P}}{d\mathbb{P}_0} \right) [X, Y] = \sum_{k=0}^D \langle h_k(Z), \zeta_k(Y) \rangle$$

676 and

$$\left\| \mathcal{P}_{\leq D} \left(\frac{d\mathbb{P}}{d\mathbb{P}_0} \right) \right\|_2^2 \leq \binom{r+D}{r}.$$

677 *Proof.* Note that the density ratio is invariant to X conditioned on Z so we can Hermite expand
678 directly in Z :

$$\mathbb{E}_0 \left[h_k(Z) \frac{d\mathbb{P}}{d\mathbb{P}_0} [X, Y] \middle| Y \right] = \mathbb{E} [h_k(Z) | Y] = \zeta_k(Y)$$

679 which implies that the Hermite coefficients of $\frac{d\mathbb{P}}{d\mathbb{P}_0}$ in Z are given by ζ_k . For the second equality, we
680 have by Lemma 14:

$$\left\| \mathcal{P}_{\leq D} \left(\frac{d\mathbb{P}}{d\mathbb{P}_0} \right) \right\|_2^2 = \sum_{k=0}^D \mathbb{E} \|\zeta_k(Y)\|_F^2 \leq \sum_{k=0}^D \binom{r+k-1}{k} = \binom{r+D}{D}.$$

681 ■

682 E Proofs of Section 5

683 *Proof of Proposition 4* We write $y = \sigma(z)$ to denote the deterministic link functions above.

684 1. Let $S = \{z \in \mathbb{R}^r; \sigma(z) = +1\}$. Let $k < r$ and consider $\mathbb{E}[h_k(z) | z \in S] = 2 \mathbb{E}[h_k(z) \mathbf{1}(z \in S)]$. Any coordinate of this tensor corresponds to a multivariate Hermite polynomial
685 $h_{\beta_1}(z_1) \dots h_{\beta_r}(z_r)$, with $\beta_1 + \dots + \beta_r = k$. Since $k < r$, there must exist a coordinate
686 j s.t. $\beta_j = 0$. By noting that $\mathbb{E}_{z_j} \mathbf{1}(z \in S) \equiv 1$ and $\mathbb{E}_{z_{-j}} [h_{\beta_1}(z_1) \dots h_{\beta_r}(z_r)] = 0$, we
687 conclude that $\mathbb{E}[h_k(z) \mathbf{1}(z \in S)] = 0$ whenever $k < r$, and analogously for $\mathbb{R}^r \setminus S$. Finally,
688 we easily verify that $\mathbb{E}[\sigma(z) z_1 z_2 \dots z_r] > 0$, which shows that $l^* = r$ and hence $k^* = r$.

689 2. This follows directly from $k^* \leq l^*$.

690 3. Define K as the intersection of the half-spaces, determined by normals v_1, \dots, v_M . From
691 the assumption that P is a r -dimensional multi-index model, $V = [v_1 \dots v_M]$ has rank r .
692 Any unit norm vector $u \in \text{span}(V)$ thus satisfies $\max_i |v_i \cdot u| \geq \epsilon > 0$ for some $\epsilon > 0$.
693 Let $\Sigma_K = \mathbb{E}[zz^\top | z \in K] - \mathbb{E}[z | z \in K] \mathbb{E}[z | z \in K]^\top$ be the covariance conditional on K .
694 From [Klivans et al., 2024a, Lemma B1], [Vempala, 2010, Lemma 4.7], for any u as above
695 it holds that $u^\top \Sigma_K u < 1$, which implies that $\text{span}(\Lambda_1) \cup \text{span}(\Lambda_2) = \mathbb{R}^r$.

696 4. The argument appears already in [Chen and Meka, 2020], but we reproduce it here in our
697 language for completeness.

We will use induction over the leaps. Suppose first $S = \emptyset$, and consider the level sets
 $B_\lambda = \{z; |y| \geq \lambda\}$. Since $y = \sigma(z)$ is continuous and $\lim_{r \rightarrow \infty} |\sigma(rz)| = \infty$ for any z , for
any $R > 0$ there exists λ such that B_λ does not contain the ball centered at 0 of radius R .
Thus

$$\text{Tr}(\mathbb{E}[ZZ^\top | Z \in B_\lambda]) = \mathbb{E}[\|Z\|^2 | Z \in B_\lambda] \geq R^2,$$

699 so if $R^2 > r$ we must have $\mathbb{E}[h_2(Z) | Z \in B_\lambda] \neq 0$, and hence $\Lambda_2 \neq \emptyset$.

Let us now iterate over leaps. Let S be the span of Λ_2 . We now consider the sets

$$B_{\lambda, \eta, S} = \{\bar{z}_S; |y| \geq \lambda, \|z_S\| \leq \eta\} \subset S^\perp.$$

700 By now viewing $\sigma(z) = \sigma(\bar{z}_S, z_S)$ as a polynomial in \bar{z}_S , we again argue that for any
701 $R > 0$ there exists λ such that $B_{\lambda, \eta, S}$ does not contain a ball of radius R , and therefore we
702 can identify another direction using the previous argument. Iterating this procedure until S
703 spans the whole \mathbb{R}^r shows that $k^* \leq 2$.

705 *Proof of Proposition 5* Suppose towards contradiction that $k^* > 2$. Then for any $g \in L^2_{\mathbb{P}_y}$ we have
 706 $\mathbb{E}[g(\sigma(z))H_2(z)] = 0$. Applying the coarea formula we obtain

$$0 = \int g(y) \left(\int_{\sigma^{-1}(y)} \frac{H_2(z)\gamma(z)}{\|\nabla\sigma(z)\|} d\mathcal{H}^{r-1}(z) \right) dy, \quad (20)$$

707 where \mathcal{H}^k is the k -dimensional Hausdorff measure. Since this must be true for any measurable g , we
 708 conclude that

$$L(y) := \int_{\sigma^{-1}(y)} \frac{H_2(z)\gamma(z)}{\|\nabla\sigma(z)\|} d\mathcal{H}^{r-1}(z) = 0 \quad \mathbb{P}_y - \text{a.e.} \quad (21)$$

709 We write $\sigma(z) = \sum_{R \in \mathcal{R}} (v_R^\top z + b_R) \cdot \mathbf{1}(z \in R)$, where \mathcal{R} are the different linear regions.

Case $r = 1$: Suppose first that there exists \bar{y} and $\epsilon > 0$ such that the level sets $\sigma^{-1}(u)$ contain no critical points for $u \in (\bar{y} - \epsilon, \bar{y} + \epsilon)$. The level sets $\sigma^{-1}(u)$ are discrete, and we claim that we can represent them as

$$\sigma^{-1}(u) = \{t_i + \theta_i(u - \bar{y})\}, \text{ where } \sigma^{-1}(\bar{y}) = \{t_i\}_{i \in \mathcal{I}},$$

710 and $\theta_i = 1/\sigma'(t_i) \neq 0$ have alternating sign. We thus have, for $u \in (\bar{y} - \epsilon, \bar{y} + \epsilon)$,

$$L(u) = \sum_{i \in \mathcal{I}} |\theta_i| h_2(t_i + \theta_i(u - \bar{y})) \gamma(t_i + \theta_i(u - \bar{y})). \quad (22)$$

711 Let us integrate this quantity twice now. Using the fact that $(h_{k-1}\gamma)' = -h_k\gamma$, we have

$$\bar{L}(u) := \int_{\bar{y}-\epsilon}^u L(v) dv \quad (23)$$

$$= - \sum_{i \in \mathcal{I}} \text{sign}(\theta_i) h_1(t_i + \theta_i(u - \bar{y})) \gamma(t_i + \theta_i(u - \bar{y})) + C, \quad (24)$$

$$\tilde{L}(u) := \int_{\bar{y}-\epsilon}^u \bar{L}(v) dv \quad (25)$$

$$= \sum_{i \in \mathcal{I}} \text{sign}(\theta_i) \theta_i^{-1} \gamma(t_i + \theta_i(u - \bar{y})) + C(u - \bar{y} + \epsilon) + \tilde{C} \quad (26)$$

$$= \sum_{i \in \mathcal{I}} |\theta_i|^{-1} \gamma(t_i + \theta_i(u - \bar{y})) + C(u - \bar{y} + \epsilon) + \tilde{C}. \quad (27)$$

712 From $L(u) = 0$ a.e. on $(\bar{y} \pm \epsilon)$ we have $\tilde{L}(u) = 0$ for all $u \in (\bar{y} \pm \epsilon)$, leading to

$$\sum_{i \in \mathcal{I}} |\theta_i|^{-1} \gamma(t_i + \theta_i(u - \bar{y})) = -C(u - \bar{y} + \epsilon) - \tilde{C}, \quad \forall u \in (\bar{y} \pm \epsilon). \quad (28)$$

713 Since all terms are analytic, we must have this equality for all u , which implies $C = \tilde{C} = 0$, but this
 714 is a contradiction, since the LHS is a sum of positive terms.

715 **Case $r > 1$** We can represent a piece-wise linear continuous function in terms of a simplex
 716 triangulation, and the values of the function at its vertices. Consider $M = \sup_z \{\sigma(z)\}$ the maximum
 717 of σ , attained at a discrete set of global maxima. Now, let us start decreasing the level set until we
 718 reach another vertex, to say M' . We will study the family of level sets $\sigma^{-1}(y)$ for $y \in [M', M]$. We
 719 reparametrize y as $y = M + u(M' - M)$, so this family can now be indexed with $u \in [0, 1]$.

For $\theta \in \mathbb{S}^{r-1}$ and $t \in \mathbb{R}$, let $E(\theta, t) := \{z; \theta^\top z = t\}$ denote a hyperplane normal to θ and passing at distance t to the origin. We can then write

$$\sigma^{-1}(u) = \cup_{R \in \bar{\mathcal{R}}} S_R(u),$$

720 where $S_R(u) = E(v_R/\|v_R\|, \|v_R\|u + b_R) \cap R$, and where $\bar{\mathcal{R}}$ is the subset of linear regions crossed
 721 by these level sets. Note that by construction this family $\bar{\mathcal{R}}$ does not depend on u .

For $u \in (\epsilon, 1 - \epsilon)$, and for each $R \in \bar{\mathcal{R}}$, we have the following representation of the level set regions:

$$S_R(u) = \{\tilde{z} = x_R + u(z - x_R); z \in S_R(1)\}. \quad (29)$$

Here, x_R denotes a local maximum of σ , which is also a vertex of the corresponding simplex region R . Consider now $\mathcal{L}(u) := \text{Tr}\{L(u)\}$. By introducing the local change of variables $\tilde{z} = \Psi_{R,u}(z) := x_R + u(z - x_R)$ for each region, we have

$$\mathcal{L}(u) = \sum_{R \in \bar{\mathcal{R}}} \theta_R \int_{S_R(u)} (\|z\|^2 - r) \gamma(z) dz \quad (30)$$

$$= u^r \sum_R \theta_R \int_{S_R(1)} (\|x_R + u(z - x_R)\|^2 - r) \gamma(x_R + u(z - x_R)) dz, \quad (31)$$

where $\theta_R = \|\nabla \sigma(z)\|^{-1}$ for $z \in R$. Observe that \mathcal{L} is analytic in \mathbb{R} , since it is a linear combination of products of analytic functions. By assumption we have that $\mathcal{L}(u) = 0$ for $u \in (0, 1)$, which implies that \mathcal{L} should vanish everywhere. But for u sufficiently large, observe that $\mathcal{L}(u)$ is a sum of strictly positive terms, which is a contradiction.

730

Proof of Proposition 6 Let $\sigma(z) = \sum_j a_j \rho(z_j)$. By definition, we have that (i) for any $\mathcal{T} : \mathbb{R} \rightarrow \mathbb{R}$ and any polynomial q of degree $< k^*(\rho)$, $\mathbb{E}[\mathcal{T}(\rho(z_j))q(z_j)] = 0$, and (ii) there exist a transformation $\zeta(y)$ such that $\mathbb{E}[\zeta(\rho(z_j))h_{k^*(\rho)}(z_j)] \neq 0$.

Let us first show that $k^* \geq k^*(\rho)$. Suppose towards contradiction that we had a measurable \mathcal{U} and multi-indices $(\beta_1, \dots, \beta_r)$ with $|\beta| < k^*(\rho)$ such that $\mathbb{E}[\mathcal{U}(\sigma(z))H_\beta(z)] \neq 0$. Then, denoting $H_\beta(z) = H_{\beta_{-j}}(z_{-j})h_{\beta_j}(z_j)$, we have

$$\mathbb{E}_{z_j} [\mathbb{E}_{z_{-j}} (\mathcal{U}(\sigma(z))H_{\beta_{-j}}(z_{-j})) h_{\beta_j}(z_j)] \neq 0 \quad (32)$$

$$\mathbb{E}_{z_j} \left[\mathbb{E}_{z_{-j}} \left\{ \mathcal{U} \left(a_j y_j + \sum_{j' \neq j} a_{j'} \rho(z_{j'}) \right) H_{\beta_{-j}}(z_{-j}) \right\} h_{\beta_j}(z_j) \right] \neq 0 \quad (33)$$

$$\mathbb{E}_{z_j} [\tilde{\mathcal{T}}_j(y_j)h_{\beta_j}(z_j)] \neq 0, \quad (34)$$

where we defined the label transformation $\tilde{\mathcal{T}}_j(y) := \mathbb{E}_z [\mathcal{U}(a_j y + \sum_{j' \neq j} a_{j'} \rho(z_{j'}))]$. We have thus reached a contradiction. Observe that the same argument also applies if one replaces $\sigma(z)$ by $\bar{\sigma}(z) = F(\rho(z_1), \dots, \rho(z_r))$ for arbitrary F .

Let us now show $k^* \leq k^*(\rho)$. By definition, we have that $\rho \in L^2(\gamma)$. Assume first that $\rho(z)$ is subexponential, and assume w.l.o.g that $a_1 = a_2 = \dots = a_r = 1$. By picking $\beta = \beta_1 = (k^*, 0, \dots, 0)$, we need to find a measurable function \mathcal{U} such that

$$\left| \mathbb{E}_y \left(\zeta(y) \int \mathcal{U}(a_1 y + \sum_{j' > 1} a_{j'} \rho(z_{j'-1})) d\gamma_{r-1}(z) \right) \right| \geq \delta > 0. \quad (35)$$

Assume first that the law of $\rho(z)$ admits a density wrt Lebesgue, denoted by $p_y(t)$. By denoting $W = a_1^{-1} \sum_{j' > 1} a_{j'} \rho(z_{j'-1})$ and $q(t)$ its density, we can rewrite the above problem as finding \mathcal{U} such that

$$|\langle \zeta, \mathcal{U} * q \rangle_{P_y}| \geq \delta > 0 \quad (36)$$

for some $\delta > 0$. We write

$$\begin{aligned} \langle \zeta, \mathcal{U} * q \rangle_{P_y} &= \langle \zeta \cdot p_y, \mathcal{U} * q \rangle_{L^2(\mathbb{R})} \\ &= \langle \hat{\zeta} * \hat{p}_y, \hat{\mathcal{U}} \cdot \hat{q} \rangle_{L^2(\mathbb{R})}, \end{aligned} \quad (37)$$

where \hat{f} denotes the Fourier transform, and where the last equality used the Parseval identity. Since $a_1 = a_2 = \dots = a_r = 1$, then $\hat{q} = \hat{p}_y^{r-1}$. Since $\rho(z)$ is subexponential, then its characteristic function \hat{p}_y is entire, which means that it cannot vanish on a set having accumulation points. Let

$Z = \{\omega; \hat{p}_y(\omega) = 0\}$ be the set of zeros, which has measure zero. Moreover, we have that \hat{q} is uniformly continuous, since it is the Fourier transform of an integrable functions. For $\eta > 0$, we define $Z_\eta = \{\omega; \exists \omega_0 \in Z; |\omega - \omega_0| \leq \eta\} \cup \{\omega; |\omega| \geq \eta^{-1}\}$ the ‘thickened’ zero set union with the set of high-frequencies. Since \hat{q} is uniformly continuous, for each $\eta > 0$ there exists $\varepsilon > 0$ such that $\inf_{\omega \notin Z_\eta} |\hat{q}(\omega)| \geq \varepsilon$. We can thus define

$$\hat{\mathcal{U}}_\eta(\omega) = \mathbf{1}(\omega \notin Z_\eta) \frac{(\hat{\zeta} * \hat{p})(\omega)}{\hat{q}(\omega)}.$$

747 For each $\eta > 0$, we have $|\hat{\mathcal{U}}_\eta(\omega)| \leq \varepsilon^{-1} |(\hat{\zeta} * \hat{p})(\omega)|$ and hence $\mathcal{U}_\eta \in L^2(\mathbb{R})$. We have

$$\langle \hat{\zeta} * \hat{p}_y, \hat{\mathcal{U}}_\eta \odot \hat{q} \rangle = \int_{Z_\eta} (\hat{\zeta} * \hat{p}_y)(\omega) \hat{\mathcal{U}}_\eta^*(\omega) \hat{q}^*(\omega) d\omega + \int_{Z_\eta^c} (\hat{\zeta} * \hat{p}_y)(\omega) \hat{\mathcal{U}}_\eta^*(\omega) \hat{q}^*(\omega) d\omega \quad (38)$$

$$= \int_{Z_\eta^c} |(\hat{\zeta} * \hat{p}_y)(\omega)|^2 d\omega. \quad (39)$$

748 Now, since \hat{p}_y is entire, in particular $\hat{\zeta} * \hat{p}_y$ is continuous, so $\int_{Z_\eta} |(\hat{\zeta} * \hat{p}_y)(\omega)|^2 d\omega \rightarrow \|\hat{\zeta} * \hat{p}_y\|^2 =$
 749 $\|\hat{\zeta} \cdot \sqrt{p_y}\|_{\mathbb{P}_y}^2 > 0$ as $\eta \rightarrow 0$, which implies that by choosing η small enough, the label transformation
 750 \mathcal{U}_η satisfies (36).

751 This shows that $[\Lambda_{k^*}]_{\beta_1} \neq 0$. Applying the same reasoning to z_j , $j \in [r]$ thus shows that $[\Lambda_{k^*}]_{\beta_j} \neq 0$.
 752 On the other hand, if we consider β with $|\beta| = k^*$ but β not of the form $\beta = (0, \dots, k^*, 0, \dots, 0)$,
 753 applying again (32) shows that $[\Lambda_{k^*}]_\beta = 0$, ie Λ_{k^*} is diagonal. We conclude that $\text{span}(\Lambda_{k^*}) = \mathbb{R}^r$
 754 and thus that $k^* \leq k^*(\rho)$.

755 We now extend the result to any $a \in \mathbb{R}^r$ such that $a_j \neq 0$ for all j . The only difference is that now
 756 $\hat{q}(\omega) = \prod_{j=2}^r \frac{a_1}{a_j} \hat{p}(\frac{a_1}{a_j} \omega)$. The zeros of \hat{q} are still a discrete set, of size at most r times the original
 757 set; and thus we obtain the same conclusion.

758 If the law of $\rho(z)$ does not admit a density, we use an approximation of unity: for small $\varepsilon > 0$,
 759 consider the noisy neuron $\tilde{y}|z = \rho(z) + \varepsilon\xi$, with $\xi \sim N(0, 1)$ independent of z . Now $p_{\tilde{y}}$ admits a
 760 density p_ε for any $\varepsilon > 0$. Note that the additive noise model carries over to the shallow NN : indeed,
 761 now we have $\tilde{y}|z = \sigma(z) + \varepsilon\|a\|_2\xi$, with $\xi \sim N(0, 1)$. From [Damian et al., 2024a, Theorem 5.2]
 762 we have that $k^*(\tilde{y}) = k^*(\rho)$. Applying the previous argument shows that the corresponding noisy
 763 shallow NN has $k^* \leq k^*(\rho)$. But since adding noise to the label cannot reduce the leap generative
 764 exponent, we conclude that the same must be true for the original shallow NN.

765 Finally, if $\rho(z)$ is not subexponential, it must be unbounded (with ρ continuous). Then we know
 766 that either $k^*(\rho) = 1$ or $k^*(\rho) = 2$. Suppose the latter. Since we are assuming in this case
 767 that ρ is continuous, the level sets $B_\lambda = \{z; |\sigma(z)| \geq \lambda\}$ have the following property: for any
 768 $R > 1$, there exists λ such that B_λ does not contain the L_∞ ball centered at 0 of radius R . Since
 769 $a_j \neq 0$ and $k^*(\rho) > 1$, the covariance $\Sigma_\lambda = \mathbb{E}[zz^\top | z \in B_\lambda]$ is diagonal, ie $\Sigma_\lambda = \text{diag}(C_\lambda)$, with
 770 $\max(C_\lambda)_j > R$. This means that at least one of the z_j will be in span of Λ_2 . Since conditioning
 771 on this z_j enables to remove the associated neuron, we can now iterate the argument to detect all
 772 coordinates with leaps of generative exponent at most two. The argument for $k^*(\rho) = 1$ is analogous.

773

Lemma 16 (Truncated and Fourier Label Transformations, [Damian et al., 2024a, Theorem 5.2]).
 Let $\mathbb{P} \in \mathcal{P}(\mathbb{R} \times \mathbb{R})$ with $\mathbb{P}_z = \gamma$ and let $k^* = k^*(\mathbb{P})$. Then there exists $R_0, \xi_0, \epsilon_0 > 0$ and $\delta_0 > 0$,
 and label transformations $\mathcal{T}_R, \tilde{\mathcal{T}}_\xi$ of the form $\mathcal{T}_R(y) = g(y)\mathbf{1}_{|y| \leq R}$ and $\tilde{\mathcal{T}}_\xi(y) = e^{i\xi y}$ such that

$$|\mathbb{E}_{\mathbb{P}}[\mathcal{T}_R(Y)h_{k^*}(Z)]| \geq \delta_0, \text{ and } |\mathbb{E}_{\mathbb{P}}[\tilde{\mathcal{T}}_\xi(Y)h_{k^*}(Z)]| \geq \delta_0$$

774 for $R \geq R_0$ and $|\xi - \xi_0| \leq \epsilon_0$.

775 *Proof of Proposition 7* By Proposition 6, we can reduce ourselves to the univariate case. We first
 776 verify that, given $\sigma : \mathbb{R} \rightarrow \mathbb{R}$, there exists $\epsilon > 0$ such that if

$$\|\sigma - \tilde{\sigma}\|_{\gamma_r} \leq \epsilon \quad (40)$$

777 then $k^*(\tilde{\sigma}) \leq k^*(\sigma) = k^*$.

778 From Lemma 16 we can use a sinusoid label transformation $\phi(y) = \cos(\xi y)$ for ξ that depends on σ
 779 such that $\mathbb{E}[\phi(\sigma(z))h_{k^*}(z)] = C \neq 0$.

780 It suffices to verify that $\mathbb{E}[\phi(\tilde{\sigma}(z))h_{k^*}(z)] \neq 0$. Let $a(z) = \sigma(z) - \tilde{\sigma}(z)$. Indeed, since ϕ is
 781 ξ -Lipschitz, we have

$$\forall z, \phi(\tilde{\sigma}(z)) = \phi(\sigma(z)) + \tilde{a}(z), \quad (41)$$

782 with $|\tilde{a}(z)| \leq \xi|a(z)|$. Thus

$$|\mathbb{E}[\phi(\tilde{\sigma}(z))h_{k^*}(z)] - \mathbb{E}[\phi(\sigma(z))h_{k^*}(z)]| = |\mathbb{E}[\tilde{a}(z)h_{k^*}(z)]| \quad (42)$$

$$\leq \|\tilde{a}\|_2 \leq \xi\|a\|_2 = \xi\|\sigma - \tilde{\sigma}\|_2, \quad (43)$$

783 so if $\epsilon < C/\xi$ we have $k^*(\tilde{\sigma}) \leq k^*(\sigma)$.

784 Finally, using a standard universal approximation theorem, e.g using the integral representation

$$\sigma(z) = \int_{\mathbb{R}^3} c\rho(az + b)d\nu(a, b, c) = \mathbb{E}_{(a,b,c) \sim \nu}[c\rho(az + b)] \quad (44)$$

785 for $\nu \in \mathcal{P}(\mathbb{R}^3)$, we can obtain $\tilde{\sigma}$ satisfying (40) by doing a Monte-Carlo approximation.

786 ■