Learning From Gaussian Data

Single and Multi-Index Models

Alex Damian

based on joint work with Loucas Pillaud-Vivien, Joan Bruna, and Jason Lee

Gaussian Single-Index Models

(X, Y) follow a Gaussian single-index model with hidden direction $w^\star \in S^{d-1}$ if:

$$X \sim N\left(0, I_d\right) \quad \text{and} \quad \mathbb{P}[Y | X] = \mathbb{P}[Y | Z] \quad \text{where} \quad Z := X \cdot w^\star$$

Examples:

- $Y = X \cdot w^\star +$noise (linear regression)
- $Y = |X \cdot w^\star| +$ noise (phase retrieval)
- $Y = \sigma(X \cdot w^\star) +$ noise (learning a single neuron)
- $Y = \xi \cdot (X \cdot w^\star)$ where $\xi \sim N(0,1)$ (multiplicative noise)

Main Question: How many samples (x_i, y_i) do you need to efficiently recover $w^\star$?

Information Theory: $n = O(d)$ samples suffice to recover $w^\star$ (maximum-likelihood)

- Naively searching for the maximum-likelihood estimator $\hat{w}_{\mathrm{MLE}}$ requires exponential time

Background: Hermite Polynomials

Orthogonal polynomials with respect to the Gaussian measure $N(0,1)$:

$$h_0(z) = 1, \quad h_1(z) = z, \quad h_2(z) = \frac{z^2 - 1}{\sqrt{2}}, \quad h_3(z) = \frac{z^3 - 3z}{\sqrt{6}}, \quad \ldots$$

Orthonormality: if $Z \sim N(0,1)$, $\mathbb{E}[h_j(Z)h_k(Z)] = \mathbf{1}_{j \neq k}$

Hermite Expansion: if $\mathbb{E}[f(Z)^2] < \infty$,

$$f(Z) = \sum_{k \geq 0} c_k h_k(Z) \quad \text{where} \quad c_k = \mathbb{E}_{Z \sim N(0,1)}[f(Z)h_k(Z)]$$

Background: Hermite Polynomials

$$x = 0\,h_0(x) + \boxed{1\,h_1(x)} + 0\,h_2(x) + 0\,h_3(x) + 0\,h_4(x) + \dots$$

$\ell^\star = 1$

$$|x| = 0.80\,h_0(x) + 0\,h_1(x) + \boxed{0.56\,h_2(x)} + 0\,h_3(x) - 0.16\,h_4(x) + \dots$$

$\ell^\star = 2$

$$x^3 - 3x = 0\,h_0(x) + 0\,h_1(x) + 0\,h_2(x) + \boxed{2.45\,h_3(x)} + 0\,h_4(x) + \dots$$

$\ell^\star = 3$

$$x^2 e^{-x^2} = 0.19\,h_0(x) + 0\,h_1(x) + 0\,h_2(x) + 0\,h_3(x) - \boxed{0.05\,h_4(x)} + \dots$$

$\ell^\star = 4$

Definition [BAGJ21]: The *information exponent* $\ell^\star$ is the first index $l \geq 1$ with non-zero Hermite coefficient c_l.

The Information Exponent

Definition [BAGJ21]**:** The *information exponent* $\ell^\star$ is the first index $l \geq 1$ with non-zero Hermite coefficient c_l.

- barrier for moment methods because it implies $\mathbb{E}[YX^{\otimes k}] = 0$ for $k < \ell^\star$
- "one-step" analyses require $n \gtrsim d^{\ell^\star}$ samples [**D**LSS22, BES+22, DKL+23, ...]
- Online SGD requires $n \gtrsim d^{1 \vee \ell^\star - 1}$ samples [BAGJ21, BBSS22, ...]
- Online SGD with smoothing requires $n \gtrsim d^{1 \vee \frac{\ell^\star}{2}}$ samples [BCR19, **D**NGL23]

Optimal Rates for Single-Index Models

Online SGD: $n \gtrsim d^{1 \vee \ell^\star - 1}$

[BAGJ21]

Smoothed SGD: $n \gtrsim d^{1 \vee \frac{\ell^\star}{2}}$

[DNGL23]

Is this optimal?

No! The information exponent is **not invariant to label transformations.**

$$g(z) = h_{10}(z)$$

$$\ell^\star(g) = 10 \qquad \ell^\star(g^2) = 2$$

$$n \gtrsim d^5 \qquad n \gtrsim d$$

Can learn with $n \gtrsim d$ samples:

1. square all the labels $y \leftarrow y^2$
2. run smoothed SGD/online SGD

[LL17, MM18, BKM+19, MLKZ20, …]

Optimal Rates for Single-Index Models

$$g(z) = h_{10}(z)$$

$$\ell^\star(g) = 10 \qquad \ell^\star(g^2) = 2$$

$$n \gtrsim d^5 \qquad n \gtrsim d$$

Can learn with $n \gtrsim d$ samples:

1. square all the labels $y \leftarrow y^2$
2. run smoothed SGD/online SGD

[LL17, MM18, BKM+19, MLKZ20, …]

Theorem [MM18]: $\exists T : \mathbb{R} \to \mathbb{R}$ such that $\ell^\star(Z, T(Y)) = 2$ if an only if:

$$\mathbb{E}[T_2(Y)^2] \neq 0 \quad \text{where} \quad T_2(Y) := \mathbb{E}[Z^2 - 1 \,|\, Y].$$

If T_2 is nonzero, $w^\star$ can be recovered with $n = O(d)$ using a spectral estimator.

The same condition on T_2 is a barrier for AMP when $n = \Theta(d)$ [BKM+19, MLKZ20]

Optimal Rates for Single-Index Models

Theorem [MM18]: $\exists T : \mathbb{R} \to \mathbb{R}$ such that $\ell^\star(Z, T(Y)) = 2$ if an only if:

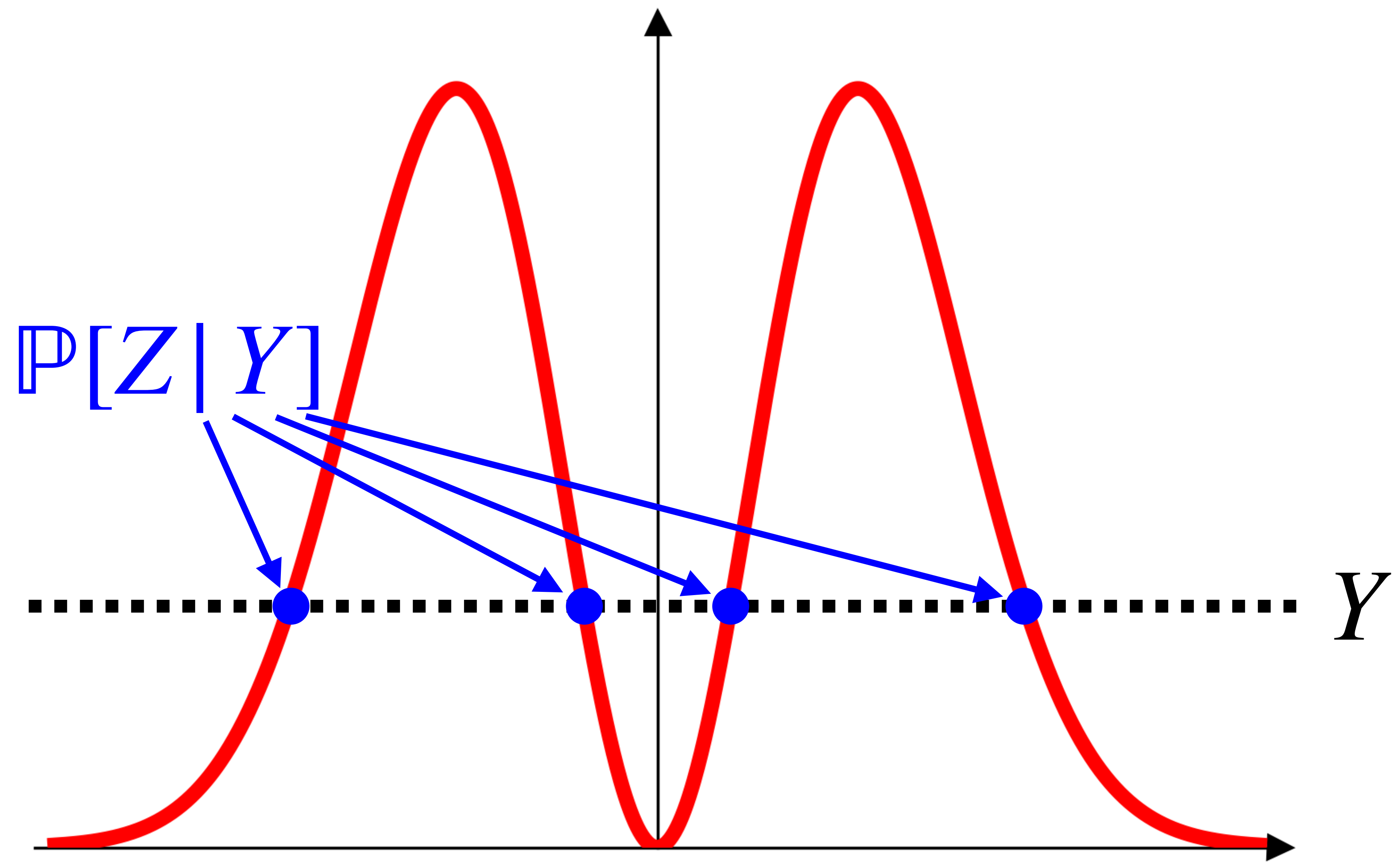
$$\mathbb{E}[T_2(Y)^2] \neq 0 \quad \text{where} \quad T_2(Y) := \mathbb{E}[Z^2 - 1 \,|\, Y].$$

If T_2 is nonzero, $w^\star$ can be recovered with $n = O(d)$ using a spectral estimator.

The same condition on T_2 is a barrier for AMP when $n = \Theta(d)$ [BKM+19, MLKZ20]

$$\mathbb{E}[Z^2 \,|\, Y] = 1 \ \forall Y$$

the bad case



$$Y = Z^2 e^{-Z^2}$$

The Generative Exponent $k^\star$ [DPVLB24]

Variational Definition: $k^\star$ is the smallest $\ell^\star$ achievable by a label transformation T:

$$k^\star := \inf_T \ \ell^\star(Z, T(Y))$$

Level Set Definition: $k^\star$ is the smallest positive integer k such that:

$$\mathbb{E}[T_k(Y)^2] \neq 0 \quad \text{where} \quad T_k(Y) := \mathbb{E}[h_k(Z) \mid Y]$$

Examples:

- All univariate polynomials have $k^\star \in \{1,2\}$ $\implies n \gtrsim d$
- $Y = Z^2 e^{-Z^2}$ has $k^\star = 4$ $\implies n \gtrsim d^2$
- For all $k \geq 1$, $\exists \sigma \in C^\infty$ such that $k^\star(\sigma) = k$ $\implies n \gtrsim d^{k/2}$

Optimal Rates for Single-Index Models

information exponent

label transformation

$$k^\star := \inf_T \ \ell^\star(Z, T(Y))$$

generative exponent

$$k^\star \le \ell^\star$$

Theorem [**D**PVLB24]

$n \gtrsim d^{\frac{k^\star}{2}} + d/\epsilon$ *samples are necessary* and sufficient to recover* $w^\star$ *to error* ϵ

Upper Bound:

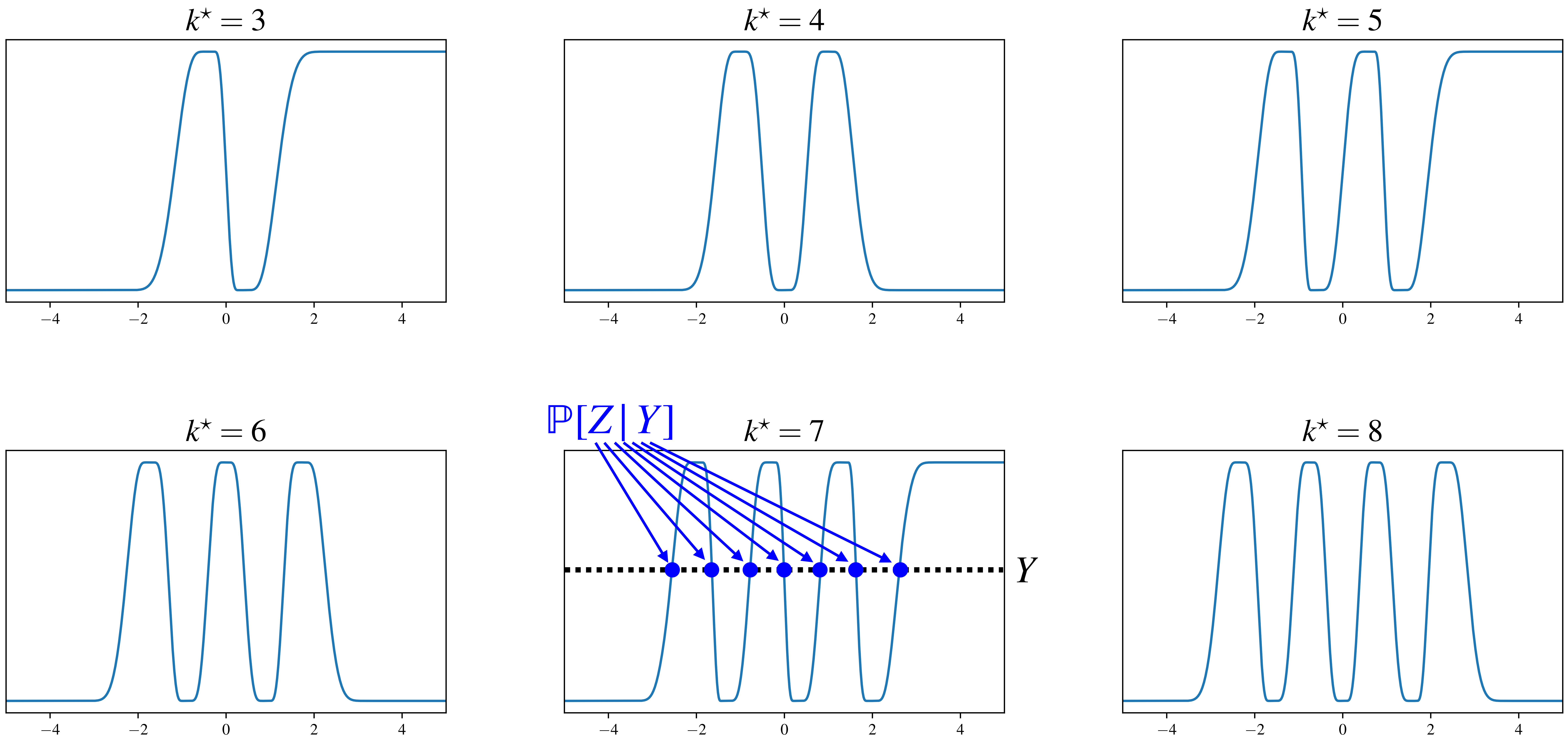
1. transform the labels $y \leftarrow T(y)$
2. run smoothed SGD [**D**NGL23]

Lower Bound:

polynomial time algorithms* cannot learn with fewer samples

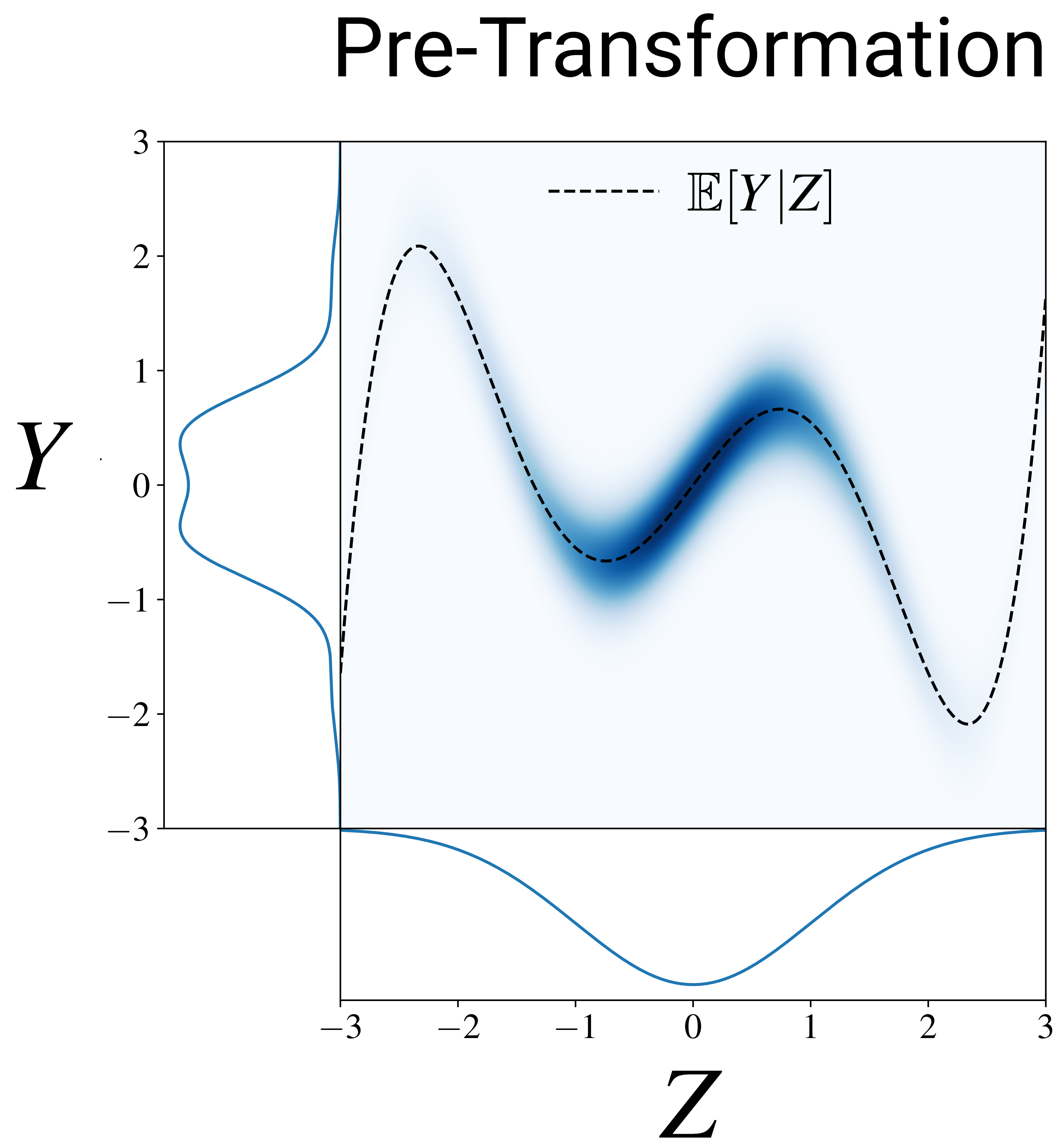
*statistical query + low degree learners

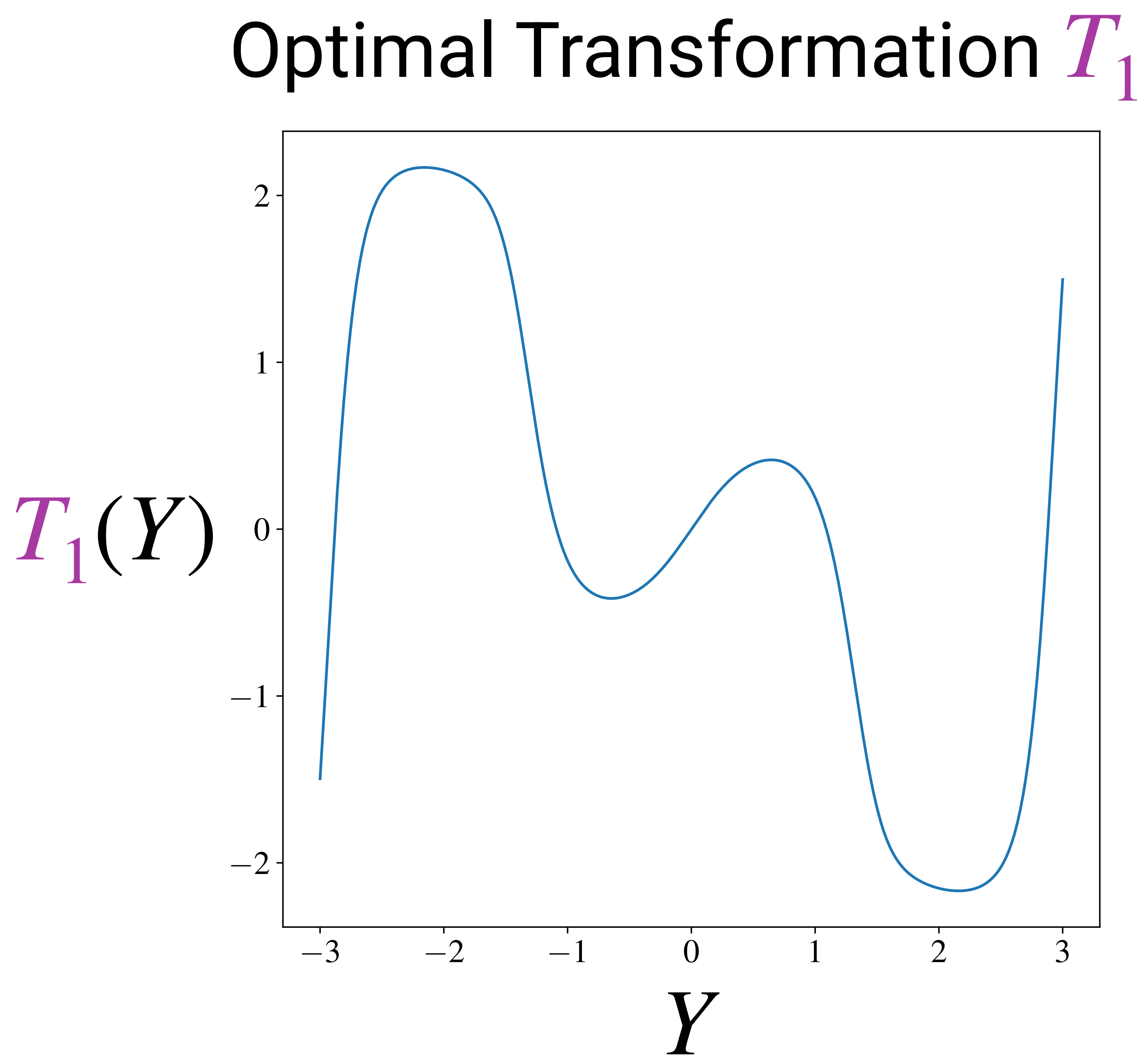
Examples of Hard Link Functions

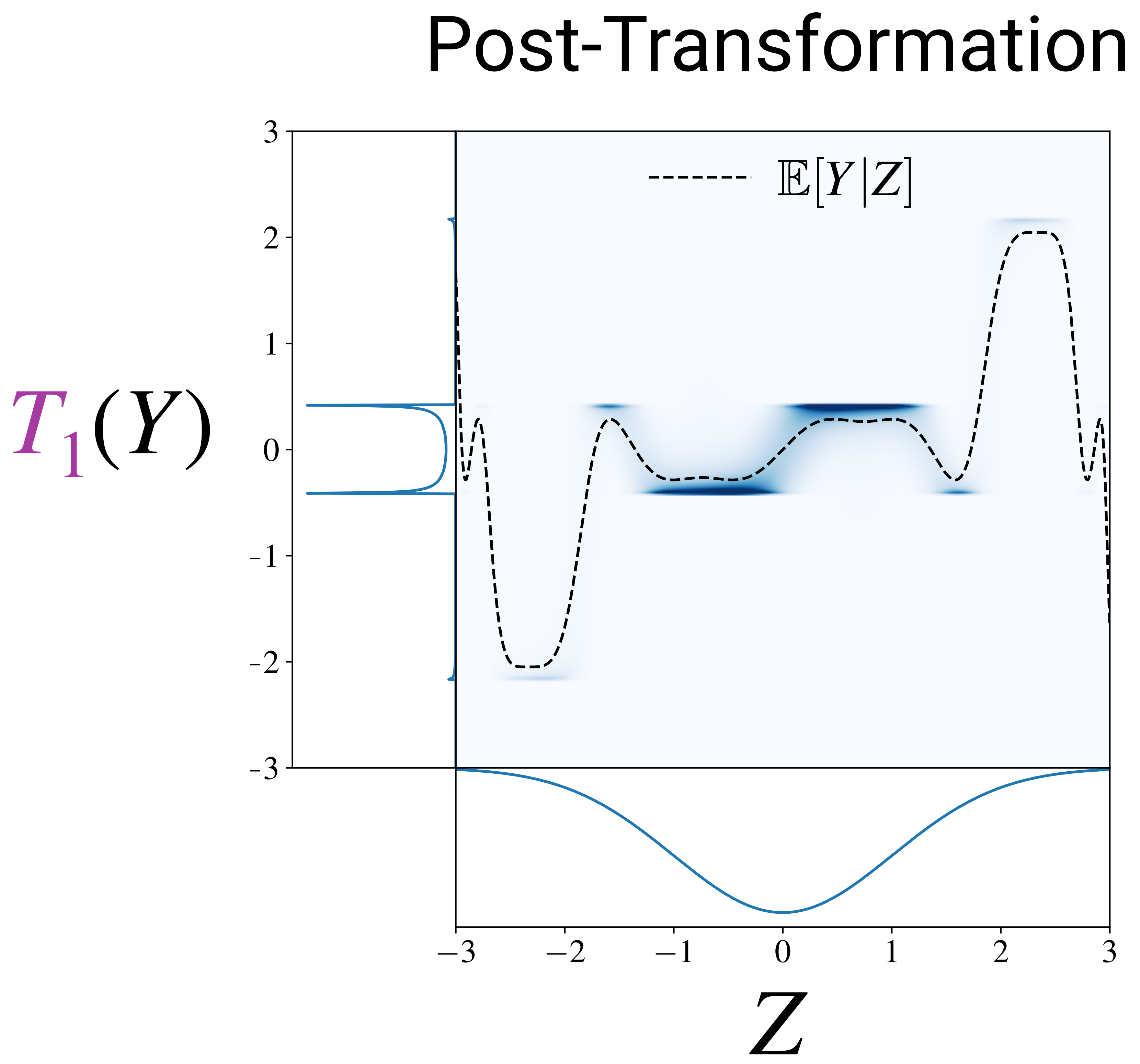


Examples of optimal transformations T

$$Y = h_5(Z) + \text{noise}, \qquad \ell^\star = 5,\ k^\star = 1$$

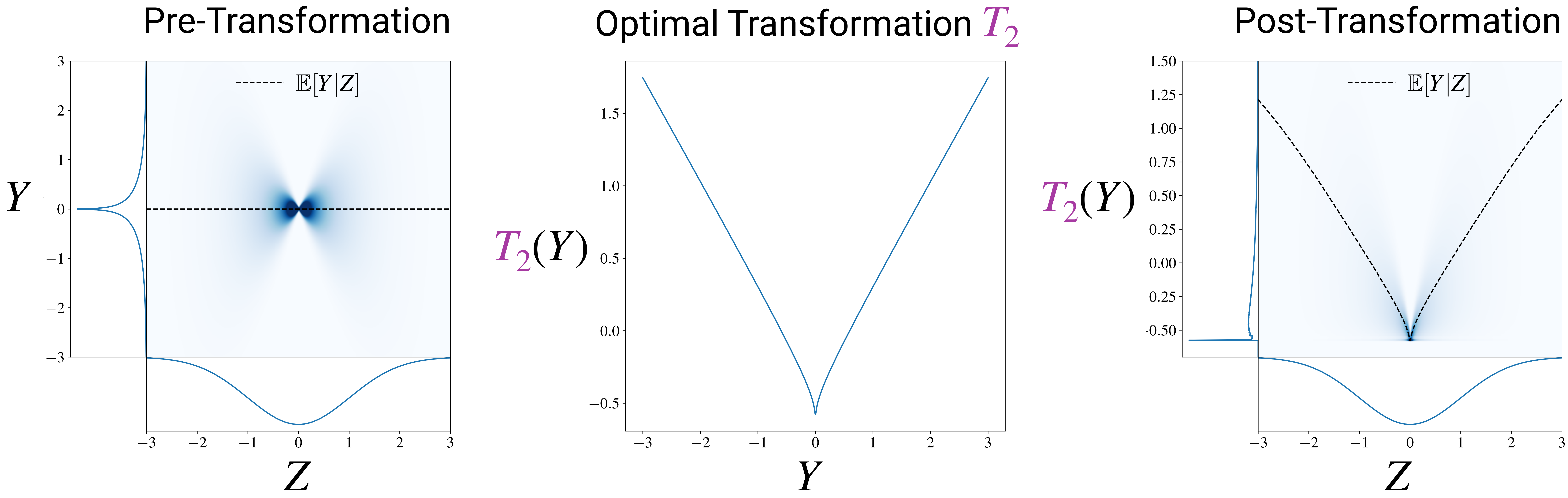






Examples of optimal transformations T

$$Y = \xi \cdot Z \text{ where } \xi \sim N(0,1), \qquad \ell^\star = \infty,\ k^\star = 2$$



Conclusion for Single-Index Models

- **Generative Exponent $k^\star$:**
  smallest information exponent $\ell^\star$ achievable by a label transformation T

- **Upper Bound:**
  For any Gaussian single-index model, $w^\star$ can be efficiently recovered to error ϵ with $n \gtrsim d^{\frac{k^\star}{2}} + d/\epsilon^2$ samples by transforming the labels

- **Lower Bound:**
  This sample complexity is tight under the statistical query (SQ) and low-degree polynomial (LDP) classes of algorithms

Gaussian Multi-Index Models

(X, Y) follow a Gaussian single-index model with hidden subspace $U^\star \subseteq \mathbb{R}^d$ if:

$$X \sim N\left(0, I_d\right) \quad \text{and} \quad \mathbb{P}[Y|X] = \mathbb{P}[Y|Z] \quad \text{where} \quad Z := \mathrm{proj}_{U^\star}(X) \in \mathbb{R}^r$$

where $r := \dim U^\star$ is the "index" or the "hidden dimension"

Examples:

- $Y = \mathrm{sign}(Z_1 \cdots Z_r)$ (parity)
- $Y = a^T \sigma(W_L \sigma(W_{l-1} \cdots \sigma(W_1 Z)))$ (deep neural network)
- $Y = \prod_j \mathbf{1}(v_j \cdot Z \geq \alpha_j)$ (intersection of halfspaces)

Information Theory: $n = O(dr)$ samples suffice to recover $U^\star$ (maximum-likelihood)

- Naively searching for the maximum-likelihood estimator $\hat{U}_{\mathrm{MLE}}$ requires exponential time

Main Question: How many samples (x_i, y_i) do you need to efficiently recover $U^\star$?

The Staircase Property

[Abbe, Boix-Adsera, Misiakiewicz 22&23]

Gaussian Parity: $Y = \mathrm{sign}(Z_1 \cdots Z_r)$

- Need to learn r directions at once (r-th order saddle)
- Gradient descent is believed to require $n \gtrsim d^{r-1}$ samples

---

Staircase Functions:

$$Y = \mathrm{sign}(Z_1) + \mathrm{sign}(Z_1 Z_2) + \ldots + \mathrm{sign}(Z_1 \cdots Z_r)$$

$k^\star = 1$, so $n = O(d)$

Next, multiply all the labels by $\mathrm{sign}(Z_1)$ and subtract 1

The Staircase Property

[Abbe, Boix-Adsera, Misiakiewicz 22&23]

Gaussian Parity: $Y = \mathrm{sign}(Z_1 \cdots Z_r)$

- Need to learn r directions at once (r-th order saddle)
- Gradient descent is believed to require $n \gtrsim d^{r-1}$ samples

---

Staircase Functions:

$$Y = \mathrm{sign}(Z_2) + \ldots + \mathrm{sign}(Z_2 \cdots Z_r)$$

$k^\star = 1$, so $n = O(d)$

Next, multiply all the labels by $\mathrm{sign}(Z_2)$ and subtract 1, …

The Staircase Property

[Abbe, Boix-Adsera, Misiakiewicz 22&23]

Gaussian Parity: $Y = \mathrm{sign}(Z_1 \cdots Z_r)$

- Need to learn r directions at once (r-th order saddle)
- Gradient descent is believed to require $n \gtrsim d^{r-1}$ samples

---

Staircase Functions:

$$Y = \mathrm{sign}(Z_r)$$

$$k^\star = 1, \text{ so } n = O(d)$$

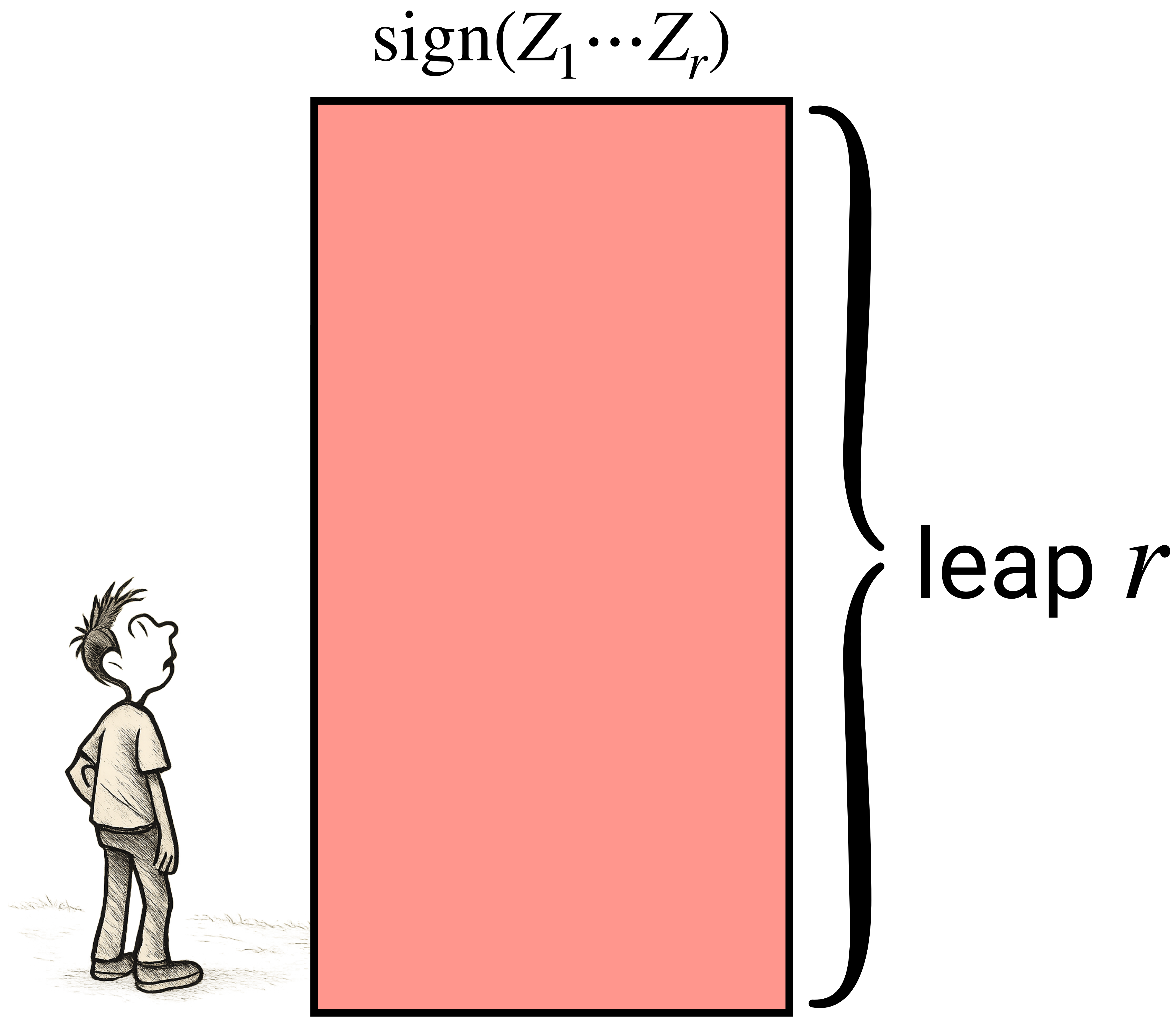
You've learned $Z_1, \ldots, Z_r$ in $O(d)$ samples!

The Staircase Property

[Abbe, Boix-Adsera, Misiakiewicz 22&23]

$Y = \text{sign}(Z_1 \cdots Z_r)$

$n = d^{r/2}$

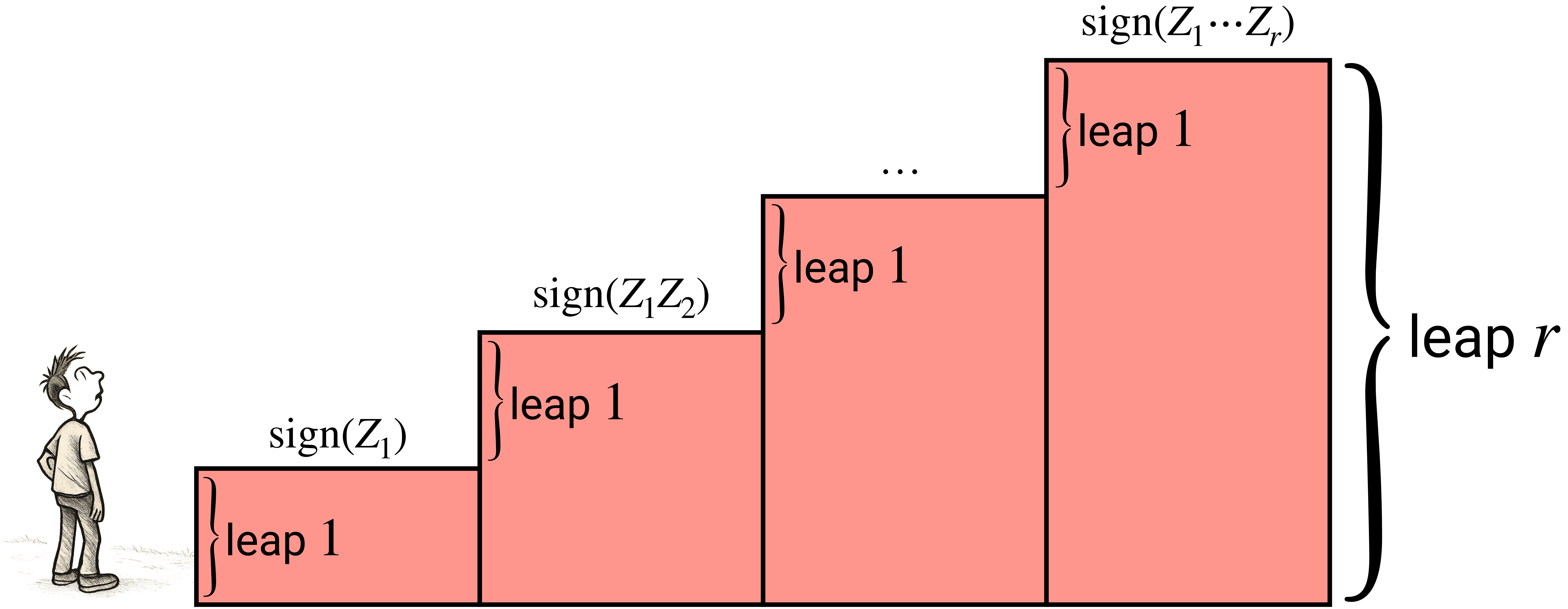


The Staircase Property

[Abbe, Boix-Adsera, Misiakiewicz 22&23]

$$Y = \mathrm{sign}(Z_1) + \cdots + \mathrm{sign}(Z_1 \cdots Z_r)$$

$n = d$

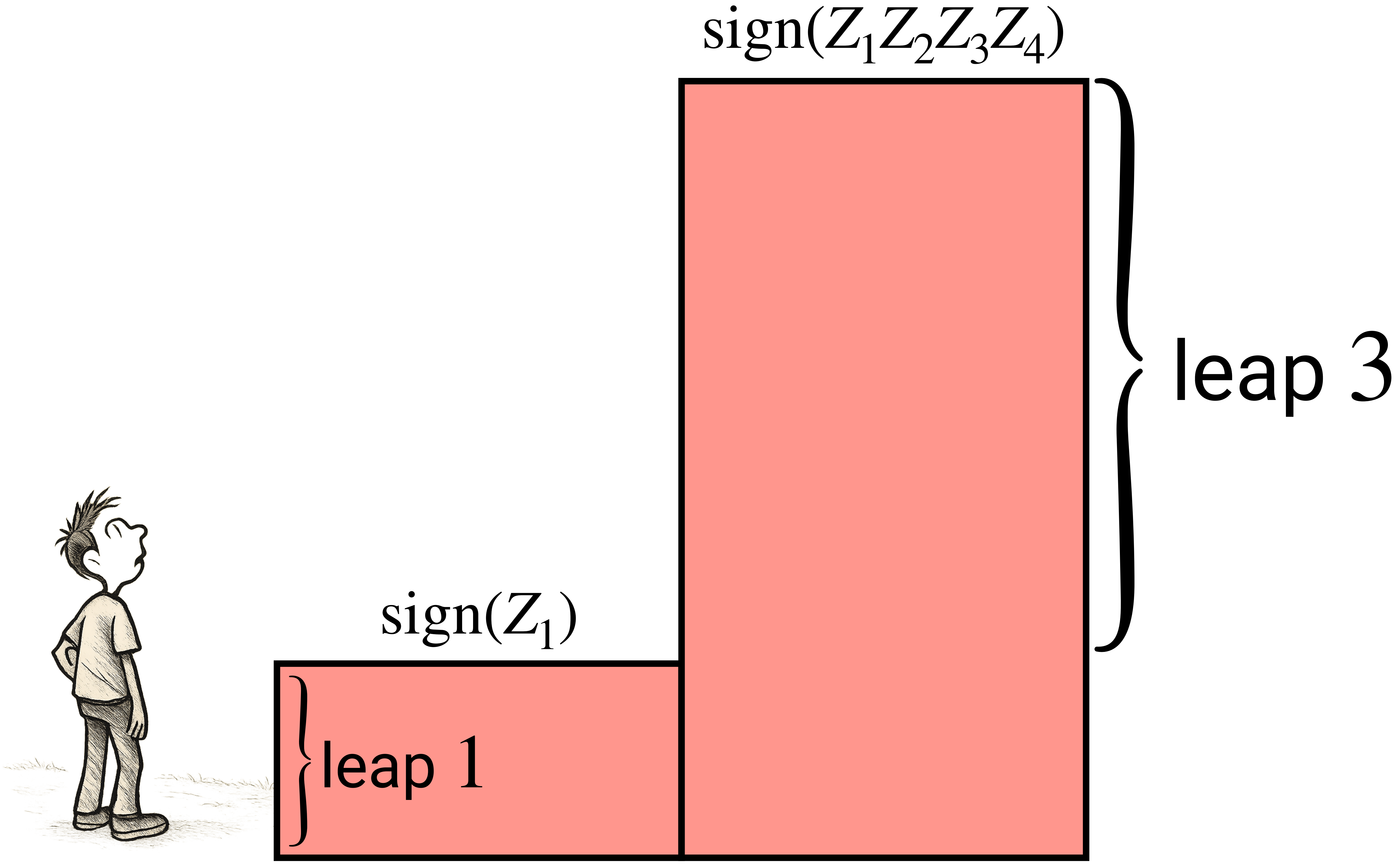


The Staircase Property

[Abbe, Boix-Adsera, Misiakiewicz 22&23]

$$Y = \mathrm{sign}(Z_1) + \mathrm{sign}(Z_1 Z_2 Z_3 Z_4)$$

$$n = d^{3/2}$$

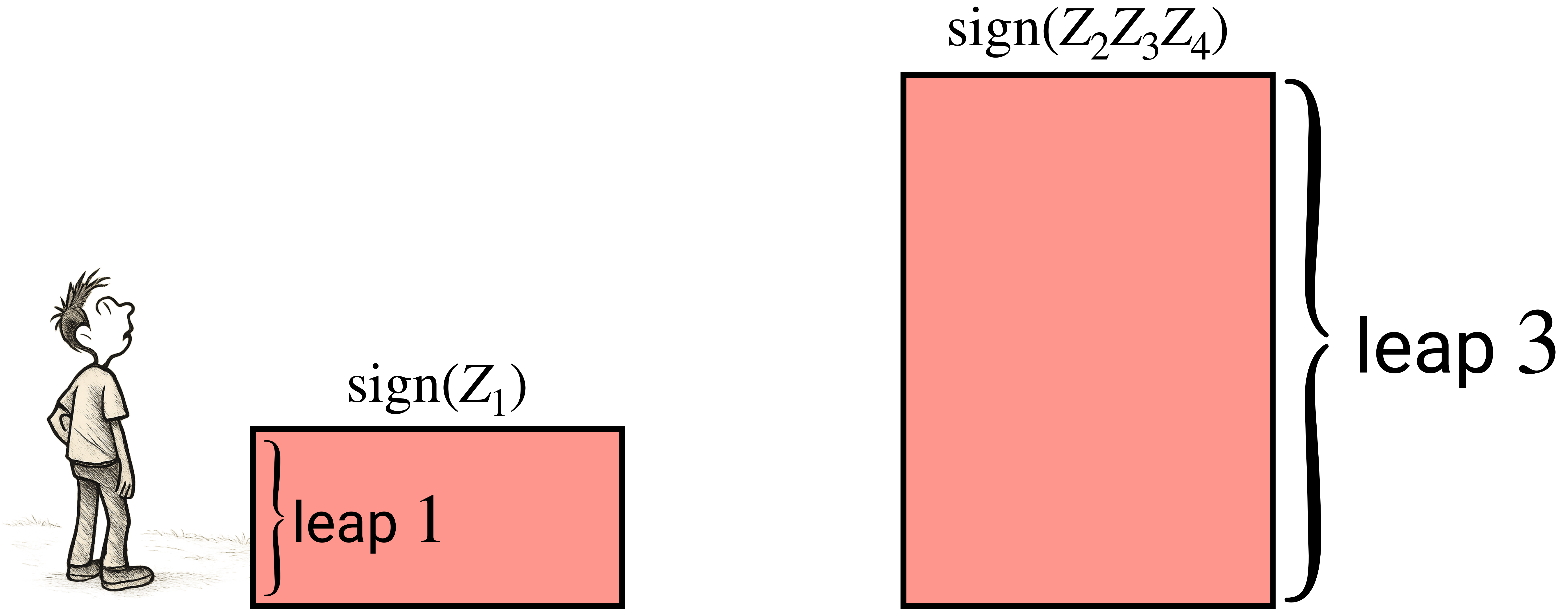


The Staircase Property

[Abbe, Boix-Adsera, Misiakiewicz 22&23]

$$Y = \mathrm{sign}(Z_1) + \mathrm{sign}(Z_2 Z_3 Z_4)$$

$$n = d^{3/2}$$

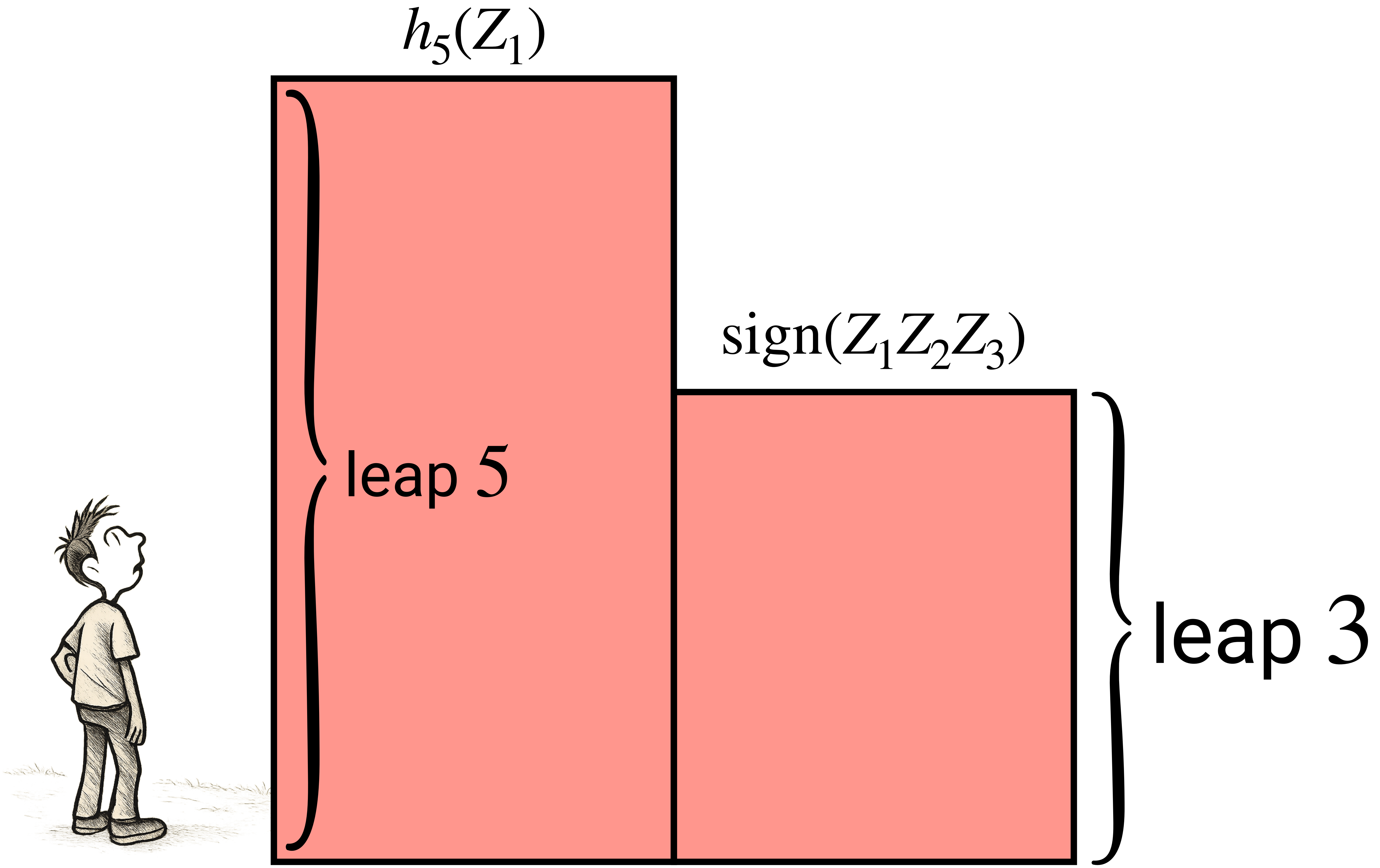


The Staircase Property + Generative Exponent

[Abbe, Boix-Adsera, Misiakiewicz 22&23] [DPVLB24]

$$Y = h_5(Z_1) + \mathrm{sign}(Z_1 Z_2 Z_3 Z_4)$$

$$n = d^{3/2}$$

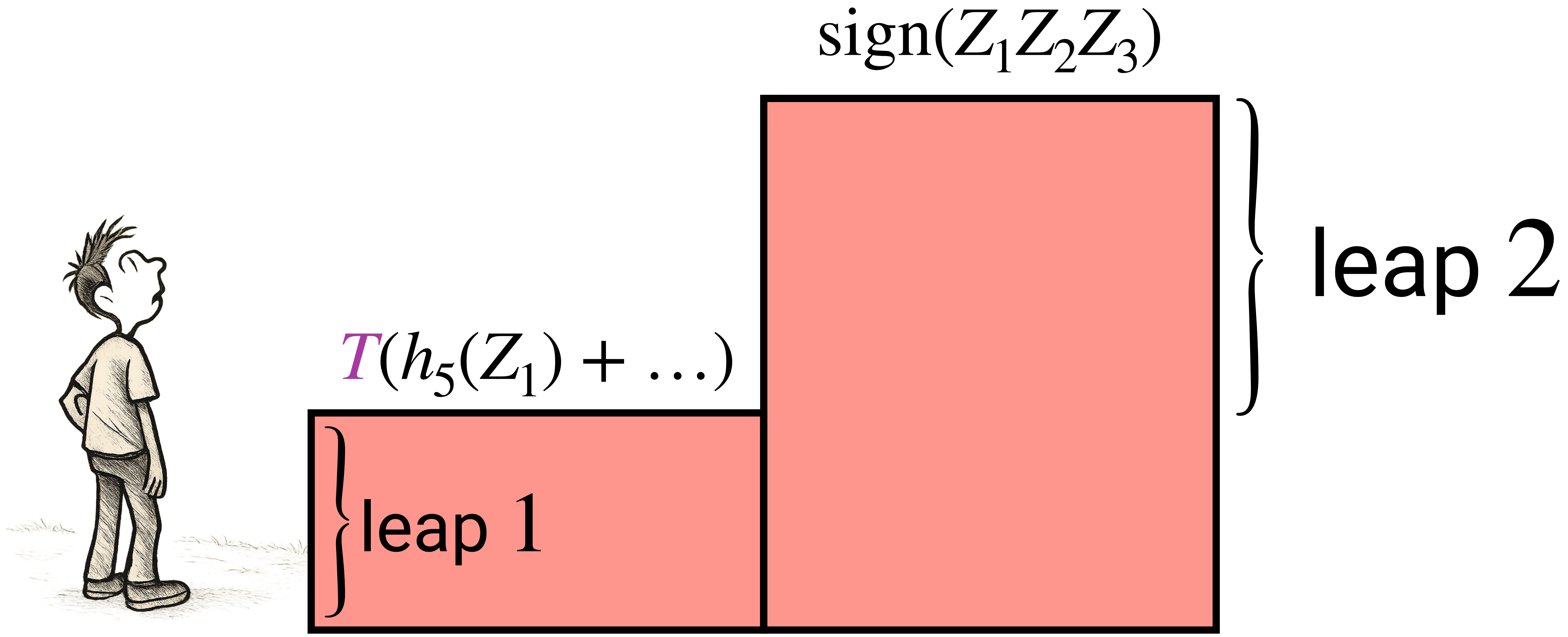


The Staircase Property + Generative Exponent

[Abbe, Boix-Adsera, Misiakiewicz 22&23] [DPVLB24]

$$T(Y) = T\Big(h_5(Z_1) + \mathrm{sign}(Z_1 Z_2 Z_3 Z_4)\Big)$$

$$n = d$$



The Staircase Property + Generative Exponent

[Abbe, Boix-Adsera, Misiakiewicz 22&23] [DPVLB24]

$$T(Y) = T\left(h_5(Z_1) + \mathrm{sign}(Z_1 Z_2 Z_3 Z_4)\right) \qquad n = d$$

“The Grand Staircase” [TDDZLK25]

Climbing the Staircase: The First Leap

$k \in \{1,2\}$ analyzed in [TDDZLK25, KZM25]

Information at leap k:

$$\bigoplus_{T} \operatorname{span}\left(\mathbb{E}\left[T(Y)\mathbf{H}_k(X)\right] \text{ reshaped as a } d \times d^{k-1} \text{ matrix} \right)$$

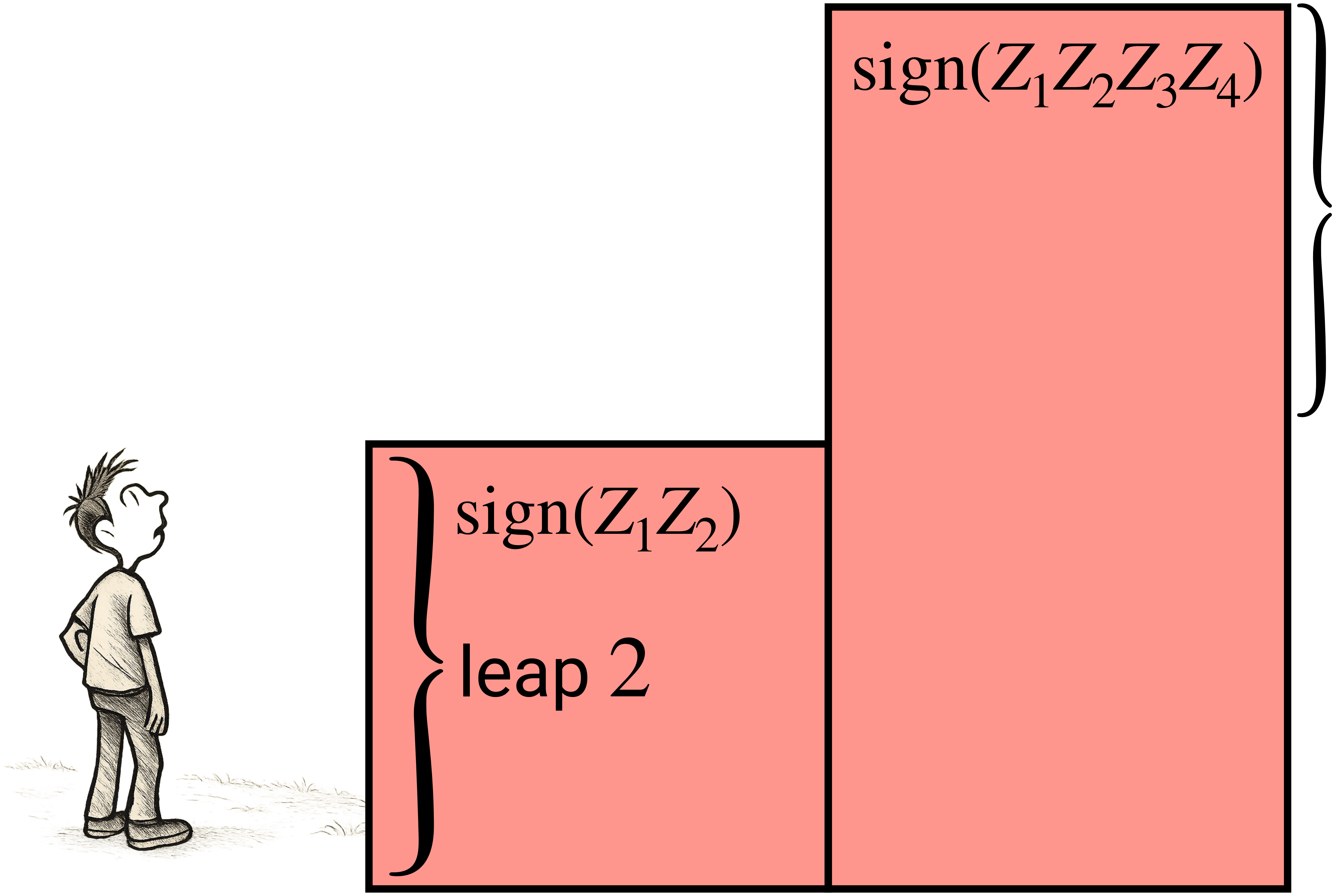
label transformation

Example: $Y = \operatorname{sign}(Z_1 Z_2) + \operatorname{sign}(Z_1 Z_2 Z_3 Z_4)$

- Information at leap 1: $\varnothing$
- Information at leap 2: $\operatorname{span}[Z_1, Z_2]$
- Information at leap 3: $\operatorname{span}[Z_1, Z_2]$
- Information at leap 4: $\operatorname{span}[Z_1, Z_2, Z_3, Z_4]$

Climbing the Staircase

$$Y = \mathrm{sign}(Z_1 Z_2) + \mathrm{sign}(Z_1 Z_2 Z_3 Z_4)$$



Need to condition on $\{Z_1, Z_2\}$ to learn $\{Z_3, Z_4\}$

Climbing the Staircase

Given: partially recovered subspace $S \subseteq \mathbb{R}^d$ (e.g. $\mathrm{span}[Z_1, Z_2]$)

Trick: just append $X_S := \mathrm{proj}_S(X)$ to your labels! $Y \leftarrow [Y, X_S] \in \mathbb{R}^{|S|+1}$

Climbing the Staircase:

$S \leftarrow \varnothing.$

While $S \neq U^\star$:

Transform both Y and X_S

$$S \leftarrow S \oplus \bigoplus_{T} \mathrm{span}\left(\mathbb{E}\left[T(Y, X_S)\mathbf{H}_k(X)\right] \text{ reshaped as a } d \times d^{k-1} \text{ matrix}\right)$$

The Generative Leap Decomposition

If we repeat $S \leftarrow S \oplus \{\text{information of order } k_i\}$, we can decompose $U^\star$ as:

$$\varnothing = S_0 \subset S_1 \subset \cdots \subset S_L = U^\star$$

such that learning S_{i+1} given knowledge of S_i is a leap of size k_i.

We define the **generative leap** to be $k^\star := \max_i k_i$ (the biggest leap) [DIKR25, **D**LB25]

Theorem [**D**LB25]: $n \gtrsim d^{1 \vee \frac{k^\star}{2}}$ is necessary* and sufficient to recover $U^\star$

Upper Bound:

1. Use a spectral method to take one step (learn S_{i+1} from S_i)
2. Iterate to climb the staircase

Lower Bound:

polynomial time algorithms* cannot learn with fewer samples

*statistical query + low degree learners

Our Estimator: A Spectral U-Statistic

Goal: estimate $\bigoplus_{T} \text{span}\left(\mathbb{E}\left[T(Y)\mathbf{H}_k(X)\right] \text{ reshaped as a } d \times d^{k-1} \text{ matrix}\right)$

Plug in estimator:

$$\text{SVD}\left[\frac{1}{n}\sum_{i=1}^{n} T(y_i)\mathbf{H}_k(X) \text{ reshaped as a } d \times d^{k-1} \text{ matrix}\right]$$

- Suffers from poor concentration 😞
- Fixable by unfolding & keeping only the non-diagonal terms (U-statistic)

$$\text{SVD}\left[\frac{1}{n(n-1)}\sum_{i \neq j} T(y_i)T(y_j)\, x_i x_j^T (x_i \cdot x_j)^{k-1}\right]$$

Our Estimator: A Spectral U-Statistic

Goal: estimate $\bigoplus_{T} \mathrm{span}\left(\mathbb{E}\left[T(Y)\mathbf{H}_k(X)\right] \text{ reshaped as a } d \times d^{k-1} \text{ matrix}\right)$

Plug in estimator:

$$\mathrm{SVD}\left[\frac{1}{n}\sum_{i=1}^{n} T(y_i)\mathbf{H}_k(X) \text{ reshaped as a } d \times d^{k-1} \text{ matrix}\right]$$

- Suffers from poor concentration 😞
- Fixable by unfolding & keeping only the non-diagonal terms (U-statistic)
- Replace $T(y_i)T(y_j)$ by a kernel $K(y_i, y_j) \Rightarrow$ “averages infinite label transformations”

$$\mathrm{SVD}\left[\frac{1}{n(n-1)}\sum_{i\neq j} K(y_i, y_j)\, x_i x_j^T (x_i \cdot x_j)^{k-1}\right]$$

Our Estimator: Climbing the Staircase

$S \leftarrow \emptyset$

for $i = 1, \ldots, m$:

Draw $\lfloor n/m \rfloor$ fresh samples

Compute the matrix U-statistic $\widehat{M}$ on $(X, \tilde{Y})$ where $\tilde{Y} = [Y | \mathrm{proj}_S(X)]$

$[\Lambda, V] \leftarrow \mathrm{SVD}(\widehat{M})$

$S \leftarrow S \oplus \mathrm{span}[v_1, \ldots, v_s]$

return S

Computing The Generative Leap $k^\star$

- **Single Index:** the generative leap and generative exponent coincide
- **Polynomials:** $k^\star \in \{1,2\}$ [CM20]
- **Gaussian parity:** $Y = \mathrm{sign}(Z_1 \cdots Z_r)$ has $k^\star = r$

  $\Rightarrow$ our upper bound gives the first algorithm that succeeds with $n = O(d^{\frac{r}{2}})$ samples
- **Intersections of halfspaces:** $k^\star \in \{1,2\}$ [Vem10]
- **Piecewise linear:** $k^\star \in \{1,2\}$

  $\Rightarrow$ implies learnability of any constant depth/width ReLU network with $n = O(d)$ samples

  $\Rightarrow$ improves a prior result of [CKM22] by allowing biases in the network

Conclusion for Multi-Index Models

- We introduced the **generative leap** $k^\star$ as a natural generalization of the generative exponent to multi-index models
- We proved an upper bound showing that for any Gaussian multi-index model, $w^\star$ can be recovered with $n \gtrsim d^{1 \vee \frac{k^\star}{2}}$ samples
- We proved this sample complexity is tight under the statistical query (SQ) and low-degree polynomial (LDP) classes
- We showed that many multi-index models, including ReLU networks, have generative leap $k^\star \in \{1,2\}$ and can be learned with $n = O(d)$ samples