Near-Optimal Statistical Query Lower Bounds for Agnostically Learning Intersections of Halfspaces with Gaussian Marginals

Clayton Sanford
Columbia Computer Science
February 14th, 2022


Me
Rocco Servedio
Daniel Hsu


Emmanouil-Vasileios Vlatakis-Gkaragkounis

Near-Optimal Statistical Query Lower Bounds for Agnostically Learning Intersections of Halfspaces with Gaussian Marginals

- **Goal:** Learn $f : \mathbb{R}^n \to \{\pm 1\}$ from hypothesis class $\mathscr{H}$ given m samples $(x_i, y_i)_{i \in [m]} \sim \mathscr{D}$.

- From m samples, obtain $f \in \mathscr{H}$ satisfying $\Pr[h(x) \neq y] \leq \mathrm{OPT} + 0.1$ for $\mathrm{OPT} = \min_{h \in \mathscr{H}} \Pr[h(x) \neq y]$ with probability ≥ 0.9.

Near-Optimal Statistical Query Lower Bounds for Agnostically Learning Intersections of Halfspaces with Gaussian Marginals

- **Problem:** Hard to prove computational lower bounds of agnostic learning.
- **Idea:** Use a restricted model for lower bounds: queries instead of samples.
- A **statistical query** takes (g, τ) with $g : \mathbb{R}^n \times \{\pm 1\} \to [-1,1]$, tolerance $\tau > 0$ and returns $q \in \mathbb{E}[g(x, y)] \pm \tau$.
- **Goal:** Learn $f : \mathbb{R}^n \to \{-1,1\}$ using SQs.
- For $\mathrm{poly}(n)$ queries of tolerance $\tau \geq 1/\mathrm{poly}(n)$, obtain f satisfying $\Pr[f(x) \neq y] \leq \mathrm{OPT} + 0.1$ for $\mathrm{OPT} = \min_{h \in \mathscr{H}} \Pr[h(x) \neq y]$ with probability ≥ 0.9.

Near-Optimal Statistical Query Lower Bounds for Agnostically Learning Intersections of Halfspaces with Gaussian Marginals

- **Problem:** Hard to prove computational lower bounds of agnostic learning.
- **Idea:** Use a restricted model for lower bounds: queries instead of samples.
- A **statistical query** takes (g, τ) with $g : \mathbb{R}^n \times \{\pm 1\} \to [-1,1]$, tolerance $\tau > 0$ and returns $q \in \mathbb{E}[g(x, y)] \pm \tau$.
- Every SQ algorithm can be simulated by a sample-based algorithm:
  - For $m = \ln(2/\delta)/(2\tau^2)$ samples, $\frac{1}{m}\sum_{i=1}^{m} g(x_i, y_i) \in \mathbb{E}[g(x, y)] \pm \tau$ with probability $\geq 1 - \delta$.
- Many sample-based algorithms can be simulated with SQs: (e.g. gradient descent, polynomial regression)
  - But, not all! Parities are PAC-learnable, but not SQ-learnable.

Near-Optimal Statistical Query Lower Bounds for Agnostically Learning Intersections of Halfspaces with Gaussian Marginals

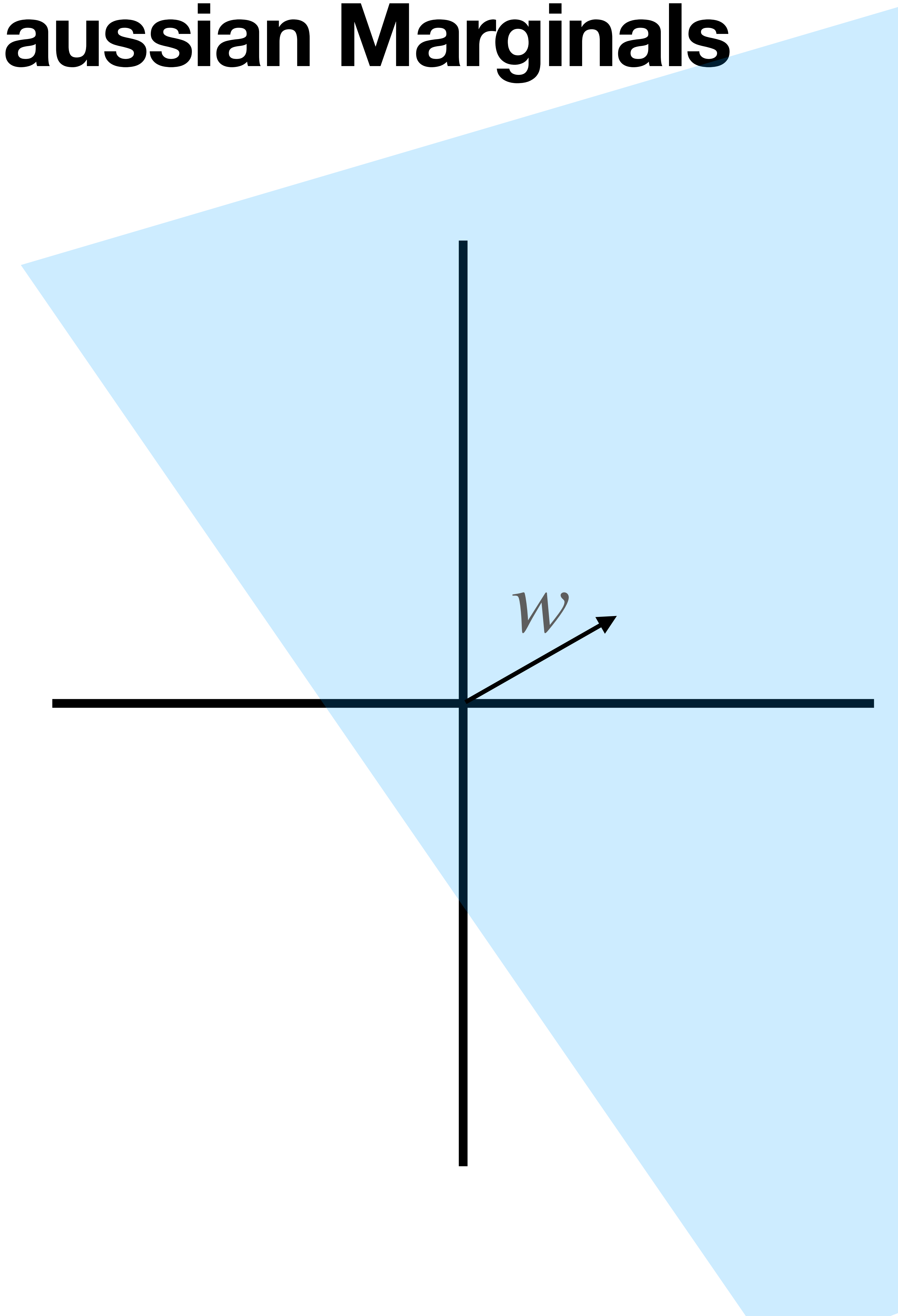
- For hypothesis class $\mathscr{H}$, show that any agnostic SQ algorithm $\mathscr{A}$ learning $\mathscr{H}$ either

    1. makes $\geq \mathrm{superpoly}(n)$ queries (🦸); or

    2. makes at least one query of tolerance $\tau \leq 1/\mathrm{superpoly}(n)$.

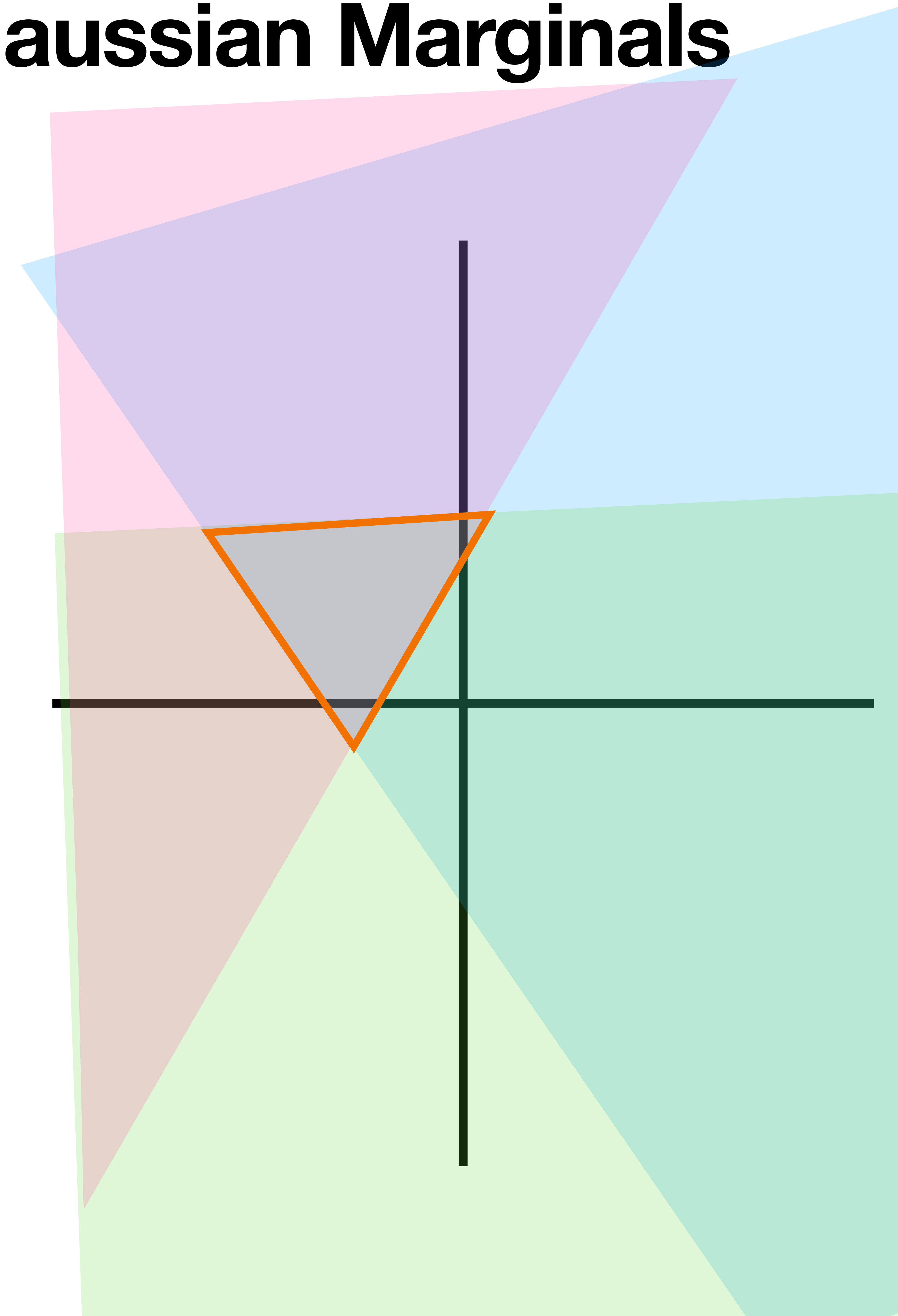
Near-Optimal Statistical Query Lower Bounds for Agnostically Learning Intersections of Halfspaces with Gaussian Marginals

- A **halfspace** is a function $h(x) = \text{sign}(w^T x - b)$.



Near-Optimal Statistical Query Lower Bounds for Agnostically Learning Intersections of Halfspaces with Gaussian Marginals

- A **halfspace** is a function $h(x) = \mathrm{sign}(w^T x - b)$.
- $\mathscr{H}_k$ is the class of all intersections of k halfspaces.
- $$f(x) = \min_{i \in [k]} \mathrm{sign}(w_i^T x - b_i) = 2\prod_{i=1}^{k} \mathbf{1}\{w_i^T x \geq b_i\} - 1$$



Near-Optimal Statistical Query Lower Bounds for Agnostically Learning Intersections of Halfspaces with Gaussian Marginals

- Features are drawn from a multivariate Gaussian distribution: $x \sim \mathcal{N}(0, I_n)$.

Near-Optimal Statistical Query Lower Bounds for Agnostically Learning Intersections of Halfspaces with Gaussian Marginals

- **[Klivans, O'Donnell, Servedio 2008]** $\mathscr{H}_k$ can be agnostically learned to accuracy ϵ with an L^1 polynomial approximation algorithm with $n^{O(\log k)}$ samples.
    - Can be implemented as an SQ algorithm that makes $n^{O(\log k)}$ queries of tolerance $n^{-O(\log k)}$.
    - **Question**: Is this dependence on k optimal?

Near-Optimal Statistical Query Lower Bounds for Agnostically Learning Intersections of Halfspaces with Gaussian Marginals (Ctd.)

- **[KOS 2008]** $\mathscr{H}_k$ can be agnostically learned to accuracy ϵ with $n^{O(\log k)}$ queries of tolerance $n^{-O(\log k)}$.

- **[Diakonikolas, Kane, Pittas, Zarifis 2021]** To agnostically learn $\mathscr{H}_k$ to accuracy ϵ, either $2^{n^{0.1}}$ queries are needed or at least one query of tolerance $\leq n^{-\tilde{\Omega}(\sqrt{\log k})}$ is necessary.

  - **Question**: Which bound has the correct dependence on k?

Near-Optimal Statistical Query Lower Bounds for Agnostically Learning Intersections of Halfspaces with Gaussian Marginals (Ctd.)

- **[KOS 2008]** $\mathscr{H}_k$ can be agnostically learned to accuracy ϵ with $n^{O(\log k)}$ queries of tolerance $n^{-O(\log k)}$.

- **[DKPZ 2021]** Requires either $2^{n^{0.1}}$ queries or at least one query of tolerance $\leq n^{-\tilde{\Omega}(\sqrt{\log k})}$.

- **Our results**: Requires either $2^{n^{0.1}}$ queries or at least one query of tolerance $\leq n^{-\tilde{\Omega}(\log k)}$.

Near-Optimal Statistical Query Lower Bounds for Agnostically Learning Intersections of Halfspaces with Gaussian Marginals (Ctd.)

[KOS '08]: $n^{O(\log k)}$

[DKPZ '21]: $n^{\tilde{\Omega}(\sqrt{\log k})}$

[HSSV '22 (?)]: $n^{\tilde{\Omega}(\log k)}$

Minimum SQ complexity needed to agnostically learn intersections of k halfspaces.

Related problems

- Realizably learning $\mathscr{H}_k$ with Gaussian marginals:
  - **[KOS ’08]** $n^{O(\log k)}$ samples (L^1 polynomial approximation)
  - **[Vempala ’10]** $\mathrm{poly}(n, k) + k^{O(\log k)}$ (PCA approach)
- Agnostically learning halfspaces ($\mathscr{H}_1$) with Gaussian marginals:
  - **[KKMS ’08]** $n^2 \log(n)/\epsilon^2$ samples
  - **[Ganzburg ’02]** + **[DKPZ ’21]** $n^{\Omega(1/\epsilon^2)}$ SQ complexity
- Learning $\mathscr{H}_k$ in distribution-free setting thought to be hard:
  - **[Klivans, Sherstov ’06]** Cryptographic hardness results
  - **[Sherstov ’13]** No efficient algos with polynomial threshold hypotheses for $\mathscr{H}_2$

Near-Optimal Statistical Query Lower Bounds for Agnostically Learning Intersections of Halfspaces with Gaussian Marginals (Ctd.)

[KOS '08]: $n^{O(\log k)}$

[DKPZ '21]: $n^{\tilde{\Omega}(\sqrt{\log k})}$

[HSSV '22 (?)]: $n^{\tilde{\Omega}(\log k)}$

Minimum SQ complexity needed to agnostically learn intersections of k halfspaces.

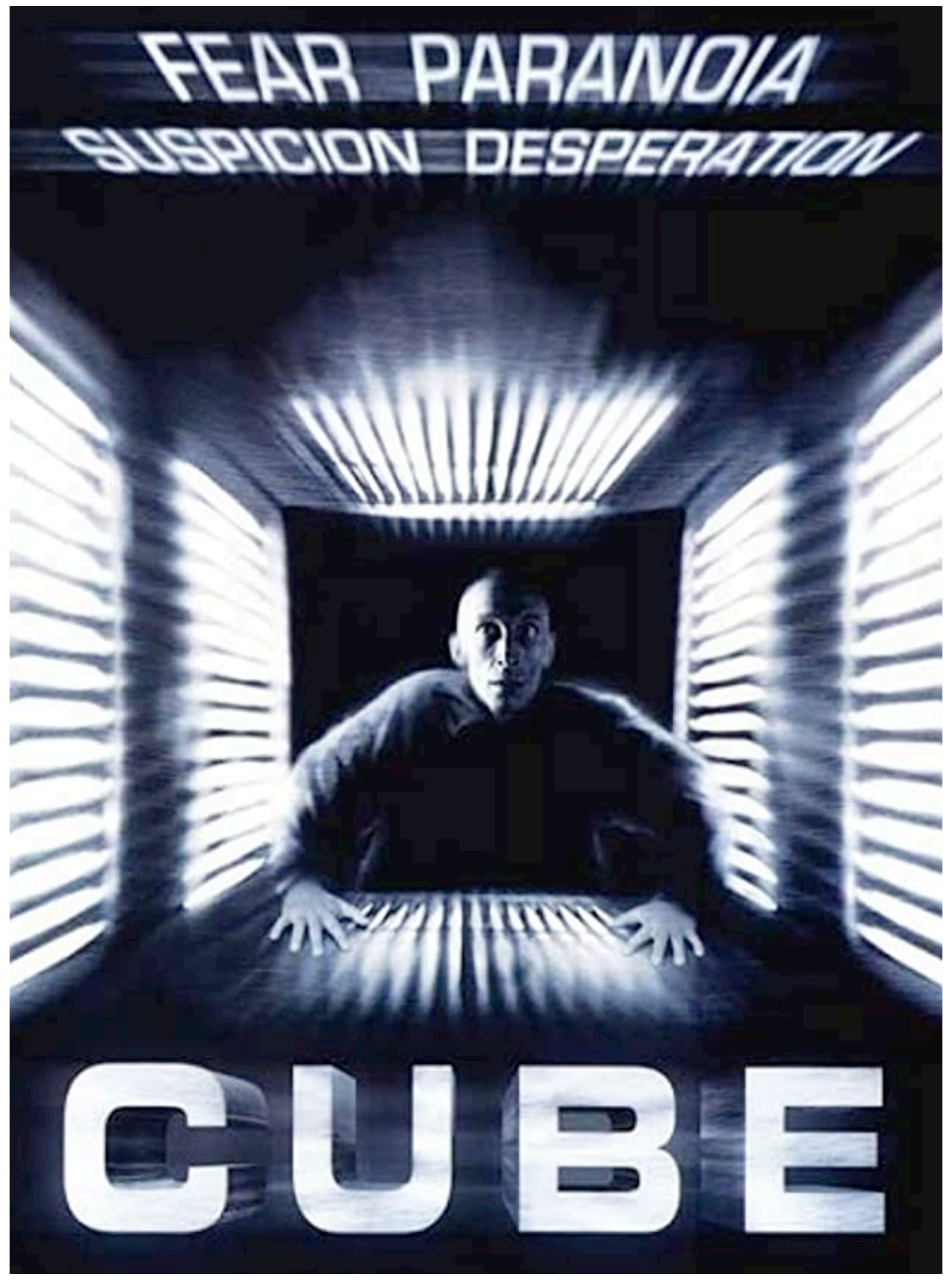
Proving an SQ lower bound

Reverse-engineering our result

- If there exists $h_1, \ldots, h_m \in \mathscr{H}$ with $|\langle h_i, h_j \rangle| = |\mathbb{E}[h_i(x)h_j(x)]| \leq 1/m$, then the **SQ-dimension** of $\mathscr{H}$ is at least m.
- Any SQ algorithm that learns $\mathscr{H}$ to error $1/2 - m^{-1/3}$ must make either $\Omega(m^{1/3})$ queries or at least one query of tolerance $O(m^{-1/3})$.
- **Intuition**: If $f : \mathbb{R}^k \to \{-1,1\}$ for $k \ll n$ cannot be approximated by low-degree polynomials, then $\mathscr{H} = \{x \mapsto f(Wx) : W \in \mathbb{R}^{k \times n}, WW^T = I\}$ has high SQ-dimension.
- Suffices to show existence of intersection of $\Theta(k)$ halfspaces f that is nearly orthogonal to low-degree polynomials.

SQ lower bounds on $\{\pm 1\}^n$
[Dachman-Soled, Feldman, Tan, Wan, Wimmer 2014]

- $f : \{\pm 1\}^k \to [-1,1]$ is d**-resilient** if for all $p \in \mathscr{P}_d$ (d-degree polys), $\langle f, p \rangle = 0$.
- f is α**-approximately** d**-resilient** if there exists $g : \{\pm 1\}^k \to [-1,1]$ such that $\|f - g\|_1 = \mathbb{E}_{x \sim \mathrm{Unif}(\{\pm 1\}^k)} |f(x) - g(x)| \le \alpha$ and g is d-resilient.
- **[DFTWW14, Thm 1.1]** For $k = n^{1/3}$, if $f : \{\pm 1\}^k \to \{\pm 1\}$ is α-approximately d-resilient, then agnostically learning $\mathscr{H} = \{f(x_S) : S \subset [n], |S| = k\}$ to excess error $(1-\alpha)/2$ requires either $n^{\Omega(d)}$ queries or at least one query of tolerance $\le n^{-\Omega(d)}$.
  - Lower bound on SQ dimension.
- **[DFTWW, Thm 1.6]** $\mathrm{Tribes} : \{\pm 1\}^k \to \{\pm 1\}$ (read-once monotone DNF) is $O(k^{-1/3})$-approximately $\Omega(\log(k)/\log\log k)$-resilient.
  - Bounds on low-degree Fourier coefficients provide transform to resilient approximation.



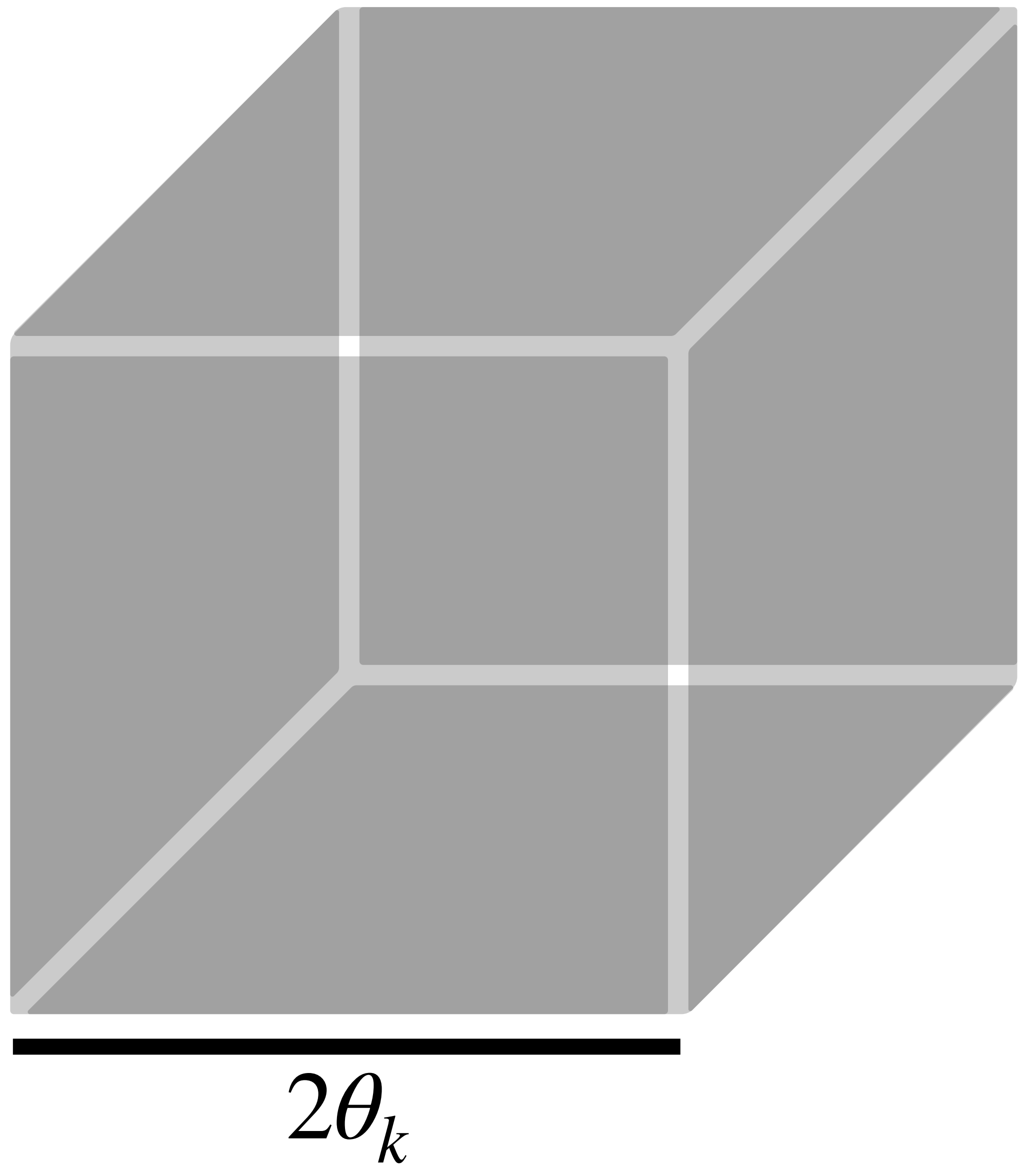
Adaptation to $\mathcal{N}(0, I_n)$
[Diakonikolas, Kane, Pittas, Zarifis 2021]

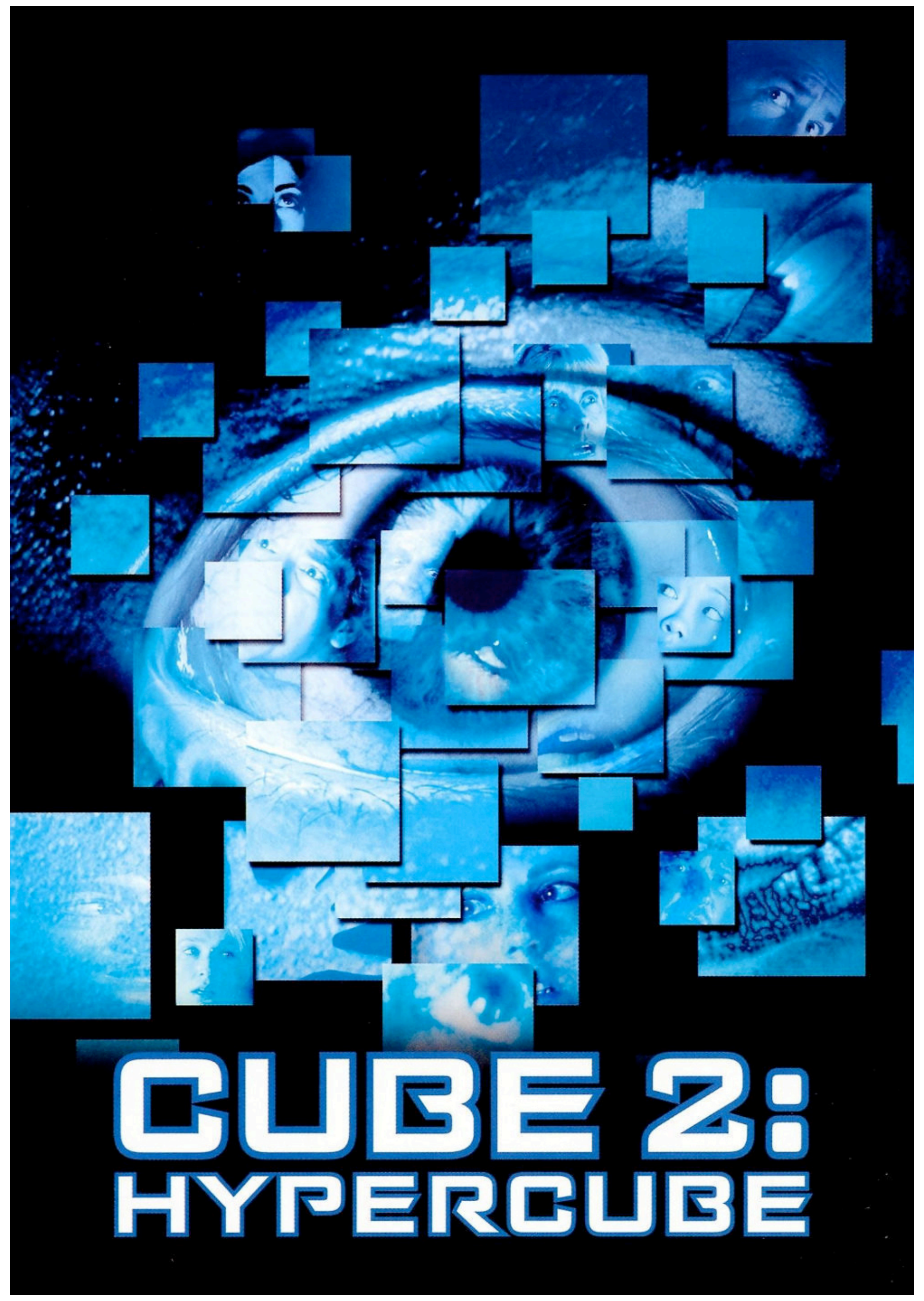
- $f : \mathbb{R}^k \to \{-1,1\}$ is **α-approximately d-resilient** if there exists $g : \mathbb{R}^k \to [-1,1]$ such that $\|f - g\|_1 \le \alpha$ and $\mathrm{Low}_d[g](x) = 0$.
- **[DKPZ '21, Prop 2.1]** f is α-approximately d-resilient iff $\|f - p\|_1 \ge 1 - \alpha$ for all $p \in \mathcal{P}_d$.
- **[DKPZ '21, Thm 1.4]** For $k = n^{0.1}$, if $\|f - p\|_1 \ge 1 - \alpha$ for all $p \in \mathcal{P}_d$, then learning $\mathcal{H} = \{x \mapsto f(Wx) : W \in \mathbb{R}^{k \times n}, WW^T = I\}$ to excess error $(1 - \alpha)/2$ requires either $2^{\Omega(n^{0.1})}$ queries or one query of tolerance $n^{-\Omega(d)}$.
- **[DKPZ '21, Thm 3.5]** For $d = \tilde{\Omega}(\sqrt{\log k})$, $\|f - p\|_1 \ge 0.1$ for all $p \in \mathcal{P}_d$ for some $f \in \mathcal{H}_k$.

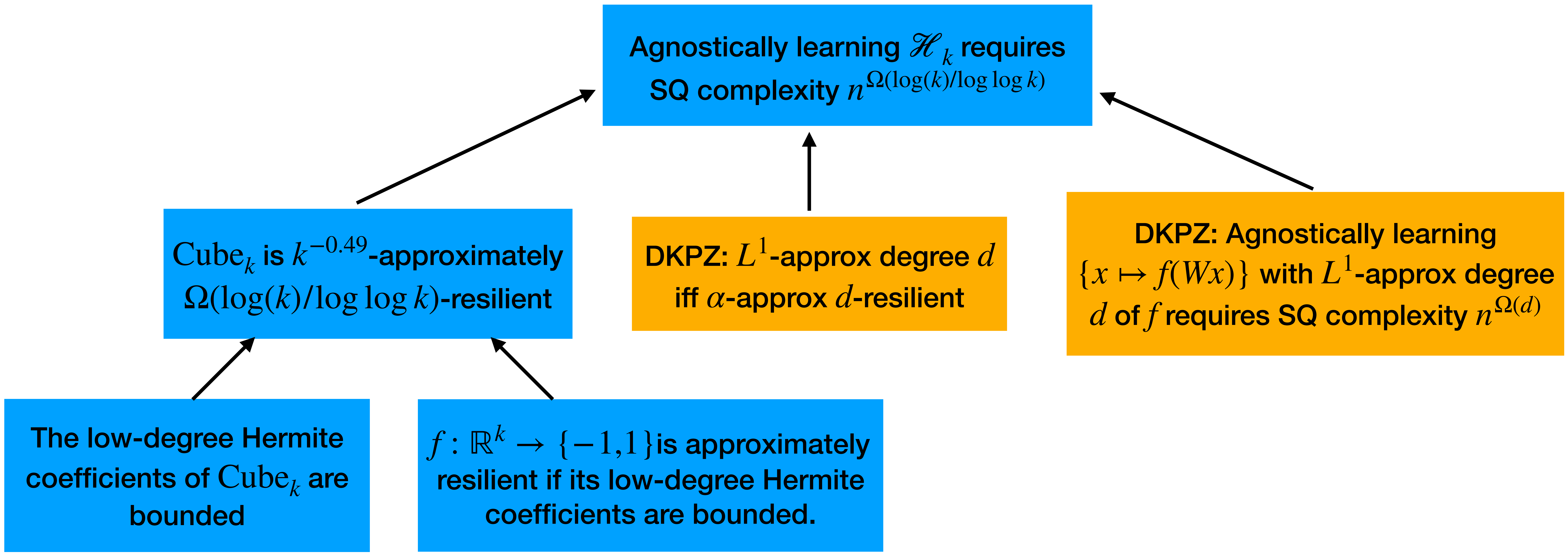
Our technical contributions:

Improving the approximate resilience bound

- $\mathrm{Cube}_k(x) = \mathrm{sign}(\theta_k - \|x\|_\infty) = 2\max_{i\in[k]} \mathbf{1}\{\,|x_i| \le \theta_k\} - 1.$
  - $\theta_k = \Theta(\sqrt{\log k})$ chosen to have $\mathbb{E}_{x\sim\mathcal{N}(0,I_k)}[\mathrm{Cube}_k(x)] = 0$.
- **Lemma:** Cube_k is $k^{-0.49}$-approximately $\Omega(\log(k)/\log\log k)$-resilient.
- **Theorem:** For $k = O(n^{0.49})$, agnostically learning $\mathscr{H} = \{x \mapsto \mathrm{Cube}_k(Wx) : W \in \mathbb{R}^{k\times n}, WW^T = I\} \subset \mathscr{H}_{2k}$ to excess error $(1-k^{-0.49})/2$ requires either $2^{O(n^{0.1})}$ queries or one query of tolerance $n^{-\Omega(\log(k)/\log\log k)}$.






Agnostically learning $\mathscr{H}_k$ requires SQ complexity $n^{\Omega(\log(k)/\log\log k)}$
Cube_k is $k^{-0.49}$-approximately $\Omega(\log(k)/\log\log k)$-resilient
DKPZ: L^1-approx degree d iff α-approx d-resilient
DKPZ: Agnostically learning $\{x \mapsto f(Wx)\}$ with L^1-approx degree d of f requires SQ complexity $n^{\Omega(d)}$
The low-degree Hermite coefficients of Cube_k are bounded
$f : \mathbb{R}^k \to \{-1,1\}$ is approximately resilient if its low-degree Hermite coefficients are bounded.

Hermite polynomials

- Multivariate probabilisist's Hermite polynomials $\{H_J\}_{J\in\mathbb{N}^k}$ is an orthogonal basis for $L^2(\mathcal{N}(0,I_k))$:
  - $\langle H_J, H_{J'}\rangle = \mathbb{E}_{x\sim\mathcal{N}(0,I_k)}[H_J(x)H_{J'}(x)] = J_1! \cdot \ldots \cdot J_k!\mathbf{1}\{J = J'\}$
- Hermite representation: $f(x) = \sum_{J\in\mathcal{N}^k} \tilde{f}(J)H_J$, for $\tilde{f}(J) = \langle f, H_J\rangle/\sqrt{J!}$
- Decomposition of f:
  - $\mathrm{Low}_d[f](x) = \sum_{|J|\leq d} \tilde{f}(J)H_J(x)$
  - $\mathrm{High}_d[f](x) = f(x) - \mathrm{Low}_d[f](x) = \sum_{|J|>d} \tilde{f}(J)H_J(x)$

Bound on low-degree Hermite coefficients of cube

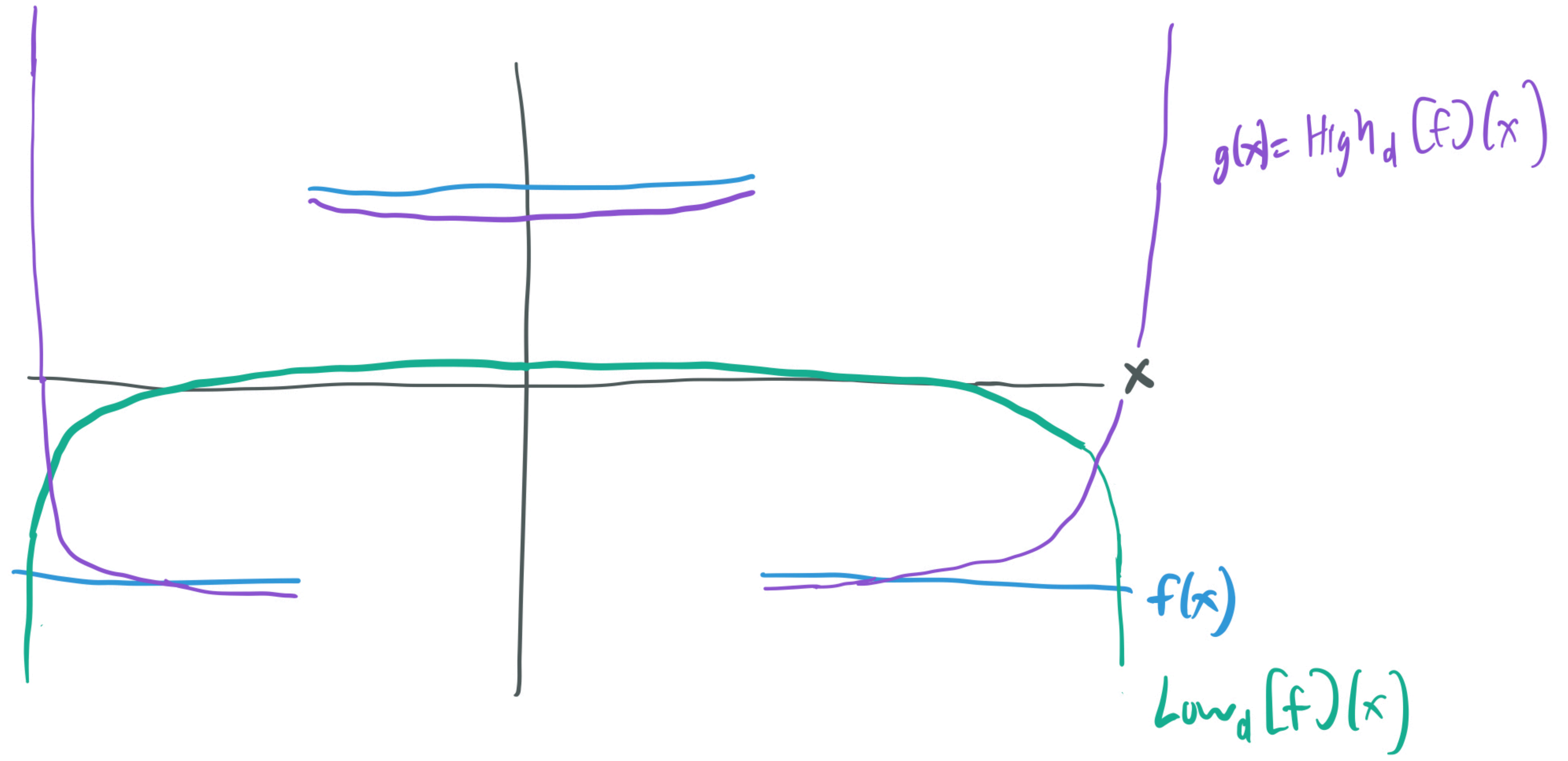
- $\mathrm{Cube}_k(x) = \mathrm{sign}(\theta_k - \|x\|_\infty)$
- **Lemma**: $\|\mathrm{Low}_d[\mathrm{Cube}_k]\|_2^2 = \sum_{|J| \le d} \widetilde{\mathrm{Cube}_k}(J)^2 \le \frac{(4 \ln k)^d}{k}$
  - For $d = \ln(k)/400 \ln \ln k$, gives $\|\mathrm{Low}_d[\mathrm{Cube}_k]\|_2^2 \le k^{-0.99}$
- Proof by exact computation of univariate Hermite coefficients of $t \mapsto \mathbf{1}\{\,|t| \le \theta_k\}$, bound by Stirling inequality and $\theta_k \in [\sqrt{2 \ln k - \ln(2 \ln k)}, \sqrt{2 \ln k}]$

Approximate Resilience: Truncation Transform

- $f : \mathbb{R}^k \to \{-1,1\}$ is **α-approximately d-resilient** if there exists $g : \mathbb{R}^k \to [-1,1]$ such that $\|f - g\|_1 \le \alpha$ and $\mathrm{Low}_d[g](x) = 0$.

- **Goal:** Transform f with small $\|\mathrm{Low}_d[f]\|_2$ to obtain **bounded** g that **approximates** f and is **uncorrelated** to low-degree polynomials.

- **Idea #1:** $g(x) := \mathrm{High}_d[f](x)$.
  - $\|g - f\|_1 \le \|\mathrm{Low}_d[f]\|_2$ 🙂
  - $\mathrm{Low}_d[g] = 0$ 😎
  - g is not bounded 💀

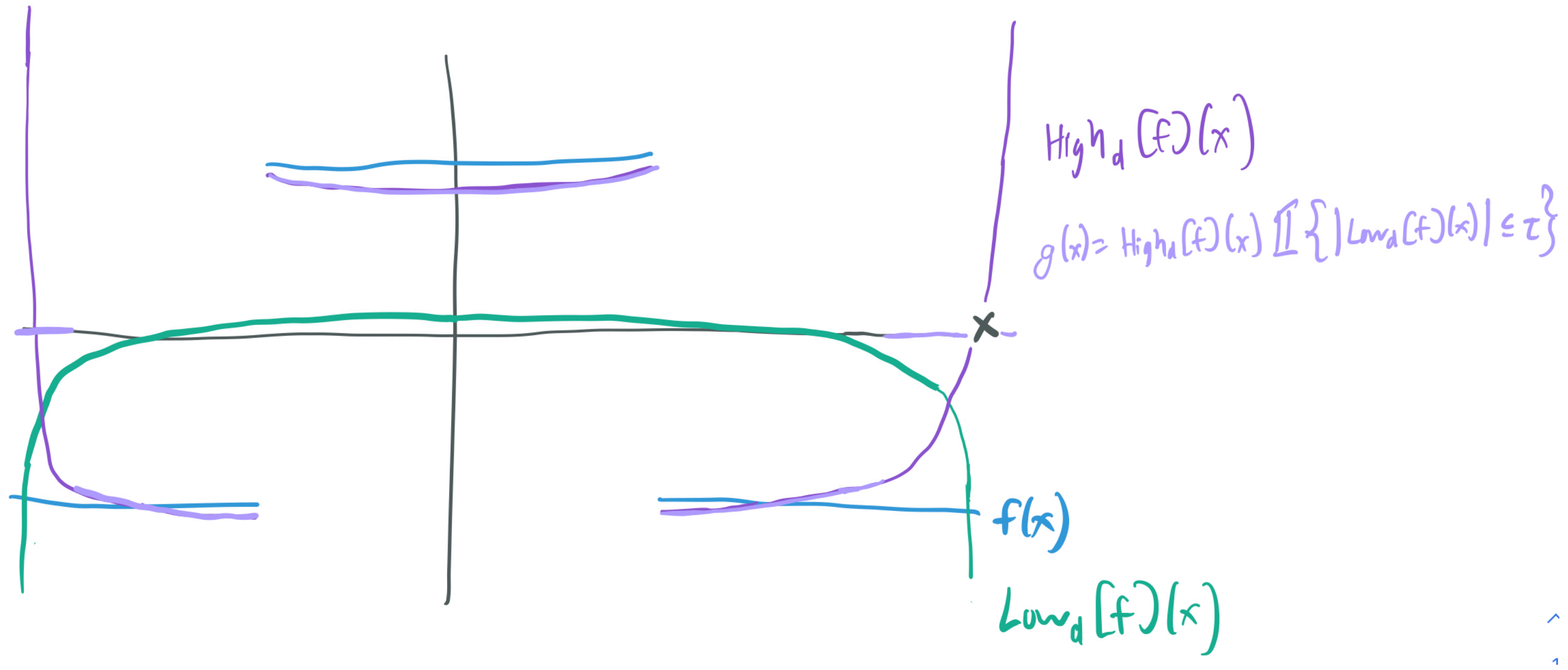


Approximate Resilience: Truncation Transform

- $f : \mathbb{R}^k \to \{-1,1\}$ is α**-approximately** d**-resilient** if there exists $g : \mathbb{R}^k \to [-1,1]$ such that $\|f - g\|_1 \le \alpha$ and $\mathrm{Low}_d[g](x) = 0$.

- **Goal:** Transform f to obtain **bounded** g that **approximates** f and is **uncorrelated** to low-degree polynomials.

- **Idea #2:** $g = \mathrm{High}_d[f](x) \cdot \mathbf{1}\{\,|\mathrm{Low}_d[f](x)| \le \eta\}$
  - For large η, $\|g - f\|_1 \le 2\|\mathrm{Low}_d[f]\|_2$ 😐
  - For large η, $\|\mathrm{Low}_d[g]\|_2 \le \|\mathrm{Low}_d[f]\|_2 / a$ 🤢
  - $\|g\|_\infty \le 1 + \eta$ 😒



Approximate Resilience: Truncation Transform

- $f : \mathbb{R}^k \to \{-1,1\}$ is α**-approximately** d**-resilient** if there exists $g : \mathbb{R}^k \to [-1,1]$ such that $\|f - g\|_1 \leq \alpha$ and $\mathrm{Low}_d[g](x) = 0$.

- **Goal:** Transform f to obtain **bounded** g that **approximates** f and is **uncorrelated** to low-degree polynomials.

- **[DFTWW ’14]**
  - $h(x) = \mathrm{High}_d[\mathrm{High}_d[f] \cdot \mathbf{1}\{\,|\mathrm{Low}_d[f]| \leq \eta\}]$
  - $g(x) = h(x)/\|h\|_\infty$.
  - 🧐

Approximate Resilience: Truncation Transform

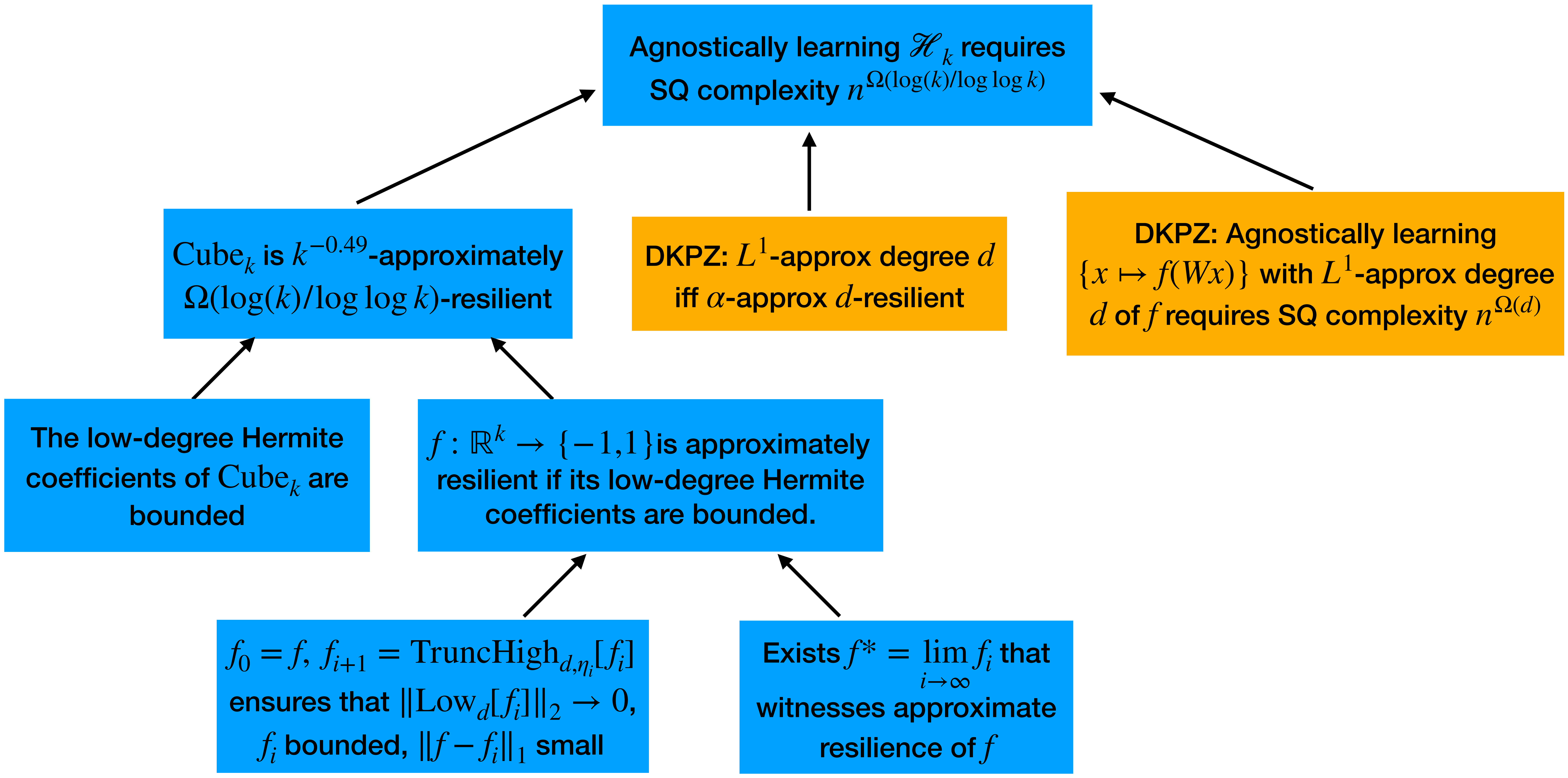
- $f : \mathbb{R}^k \to \{\pm 1\}$ is **α-approximately d-resilient** if there exists $g : \mathbb{R}^k \to [-1,1]$ such that $\|f - g\|_1 \le \alpha$ and $\mathrm{Low}_d[g](x) = 0$.

- $\mathrm{TruncHigh}_{d,\eta}[f] = \mathrm{High}_d[f](x) \cdot \mathbf{1}\{\, |\mathrm{Low}_d[f](x)| \le \eta\}$

- Let $f_0 := f$ and $f_{i+1} = \mathrm{TruncHigh}_{d,\eta_i}[f_i]$ for $i \to \infty$.

- For some decaying η_i and $\alpha = k^{0.49}$, have (1) $\|f_{i+1}\|_\infty \le \|f_i\|_\infty + \alpha/(3 \cdot 2^{i+1})$, (2) $\lim_{i\to\infty} \|\mathrm{Low}_d[f_i]\| = 0$, and (3) $\|f_{i+1} - f_i\|_1 \le \alpha/(3 \cdot 4^i)$.

- By limit argument, exists f^* with (1) $\|f^*\|_\infty \le 1 + \alpha/3$, (2) $\mathrm{Low}_d[f^*] = 0$, and (3) $\|f - f^*\|_1 \le 2\alpha/3$. Let $g := f^*/\|f\|_\infty$.



Agnostically learning $\mathscr{H}_k$ requires SQ complexity $n^{\Omega(\log(k)/\log\log k)}$
Cube_k is $k^{-0.49}$-approximately $\Omega(\log(k)/\log\log k)$-resilient
DKPZ: L^1-approx degree d iff α-approx d-resilient
DKPZ: Agnostically learning $\{x \mapsto f(Wx)\}$ with L^1-approx degree d of f requires SQ complexity $n^{\Omega(d)}$
The low-degree Hermite coefficients of Cube_k are bounded
$f : \mathbb{R}^k \to \{-1,1\}$ is approximately resilient if its low-degree Hermite coefficients are bounded.
$f_0 = f$, $f_{i+1} = \mathrm{TruncHigh}_{d,\eta_i}[f_i]$ ensures that $\|\mathrm{Low}_d[f_i]\|_2 \to 0$, f_i bounded, $\|f - f_i\|_1$ small
Exists $f^* = \lim_{i\to\infty} f_i$ that witnesses approximate resilience of f

What else is there?

- Second proof for larger $k = 2^{O(n^{0.245})}$ (rather than $k = O(n^{0.49})$) bounds on L^1 approximate degree of random intersection of halfspaces.
    - Based on hardness of weak-learning intersections of halfspaces with membership queries **[De, Servedio 2021]**
- Dependence on accuracy ϵ: $n^{\Omega(\log(k)/\log\log k + 1/\epsilon^2)}$ bound by augmenting construction with a single centered halfspace **[Ganzburg 2002]**
- Optimality of learning families with *Gaussian surface area* $\leq s$ with L^1 polynomial approximation: SQ complexity of $n^{\Omega(s^2/\log s)}$, vs $n^{O(s^2)}$ **[KOS 2008]**

Thanks!