Strong Converses for High-dimensional Statistical Estimation

Ramji Venkataramanan
University of Cambridge

Oliver Johnson
University of Bristol

Beyond IID 2018
Inference set-up

Want to estimate $\theta \in \mathcal{F}$ from data $Y = (Y_1, \ldots, Y_n)$

Data $Y$ generated according to $P_\theta(Y)$

How well can we estimate $\theta$ as the number of samples grows?
Inference set-up

Want to estimate $\theta \in \mathcal{F}$ from data $Y = (Y_1, \ldots, Y_n)$

Data $Y$ generated according to $P_\theta(Y)$

How well can we estimate $\theta$ as the number of samples grows?

**Density Estimation** [Yu ’97]

$\mathcal{F}$: smooth densities on $[0, 1]$ with bounded second derivative

For $\theta \in \mathcal{F}$, samples $Y_1, \ldots, Y_n$ drawn i.i.d. $\sim \theta$

Measure of performance:

$$d(\theta, \hat{\theta}) = \int_0^1 \left( \sqrt{\theta(x)} - \sqrt{\hat{\theta}(x)} \right)^2 \, dx$$
Compressed sensing

Vector $\theta \in \mathcal{F}$ observed through linear model:

$$y = A \theta + \text{noise}$$

$\mathcal{F}$: unit norm vectors in $\mathbb{R}^n$ with at most $k$ non-zeros

How well can we estimate $\theta$?

Measure of performance:

$$M^*(A) := \inf_{\hat{\theta}} \sup_{\theta \in \mathcal{F}} \mathbb{E} \left[ \frac{1}{n} \| \hat{\theta}(y) - \theta \|^2 \right],$$

Candès, Davenport, *How well can we estimate a sparse vector?*, 2013
Loss function and Risk

Want to estimate $\theta \in \mathcal{F}$ from data $\mathbf{Y} = (Y_1, \ldots, Y_n)$

Data $\mathbf{Y}$ generated according to $P_\theta(\mathbf{Y})$

Performance of an estimator $\hat{\theta}$ measured via $d(\theta, \hat{\theta}(\mathbf{Y}))$

Loss function $d$ is a distance or semi-distance

Risk $R(\theta, \hat{\theta}) = \mathbb{E} \left[ d(\theta, \hat{\theta}) \right]$

GOAL: Lower bounds on the minimax risk

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathcal{F}} \mathbb{E} \left[ d(\theta, \hat{\theta}) \right]$$
Standard approach (see Tsybakov 2009)

For any $\psi_n > 0$,

$$\Pr \left( d(\theta, \hat{\theta}) \geq \psi_n \right) \leq \frac{1}{\psi_n} \mathbb{E} \left[ d(\theta, \hat{\theta}) \right]$$
Standard approach (see Tsybakov 2009)

For any $\psi_n > 0$,

$$\mathbb{P} \left( d(\theta, \hat{\theta}) \geq \psi_n \right) \leq \frac{1}{\psi_n} \mathbb{E} \left[ d(\theta, \hat{\theta}) \right]$$

Hence

$$\sup_{\theta \in \mathcal{F}} \mathbb{E} \left[ d(\theta, \hat{\theta}) \right] \geq \psi_n \sup_{\theta \in \mathcal{F}} \mathbb{P} \left( d(\theta, \hat{\theta}) \geq \psi_n \right)$$

Want to choose $\psi_n$ such that prob. is bounded below by a constant
Packing set

$$\sup_{\theta \in \mathcal{F}} \mathbb{E} \left[ d(\theta, \hat{\theta}) \right] \geq \psi_n \sup_{\theta \in \mathcal{F}} \mathbb{P} \left( d(\theta, \hat{\theta}) \geq \psi_n \right)$$

Construct a packing set \( \{\theta_1, \ldots, \theta_M\} \subseteq \mathcal{F} \) such that

$$d(\theta_i, \theta_j) \geq d_{\min} = 2\psi_n, \quad \text{for all } i \neq j$$

▶ Existence of packing set can be generally guaranteed via Gilbert-Varshamov bound or the probabilistic method
Packing set

\[
\sup_{\theta \in \mathcal{F}} \mathbb{E} \left[ d(\theta, \hat{\theta}) \right] \geq \psi_n \sup_{\theta \in \mathcal{F}} \mathbb{P} \left( d(\theta, \hat{\theta}) \geq \psi_n \right)
\]

Construct a packing set \( \{\theta_1, \ldots, \theta_M\} \subseteq \mathcal{F} \) such that

\[
d(\theta_i, \theta_j) \geq d_{\text{min}} = 2\psi_n, \quad \text{for all } i \neq j
\]

- Existence of packing set can be generally guaranteed via Gilbert-Varshamov bound or the probabilistic method

- Idea: Any estimator \( \hat{\theta} \) defines an \( M \)-ary hypothesis test between \( \{\theta_1, \ldots, \theta_M\} \)

\[
\hat{i} = \arg \min_{1 \leq j \leq M} d(\theta_j, \hat{\theta})
\]
Channel coding interpretation

- Channel $P_{Y|\theta} := P_\theta(Y)$

- Codebook $\{\theta_1, \ldots, \theta_M\}$

- 'Transmitted' codeword $\theta = \theta_i$

- Channel output $Y = (Y_1, \ldots, Y_n)$

- Minimum-distance decoder Distance measured via $d(\cdot, \cdot)$
  Decode codeword that is closest to $\hat{\theta}(Y)$. 
Probability of decoding error

Minimum distance between codewords is \( d_{\text{min}} = 2\psi_n \Rightarrow \)
Decoder makes error only if \( d(\theta_i, \hat{\theta}) \geq \frac{d_{\text{min}}}{2} = \psi_n \Rightarrow \)

\[
\mathbb{P} \left( \hat{i} \neq i \mid \theta_i \text{ true codeword} \right) \leq \mathbb{P} \left( d(\theta_i, \hat{\theta}) \geq \psi_n \right)
\]
Probability of decoding error

Minimum distance between codewords is $d_{\text{min}} = 2\psi_n \Rightarrow$

Decoder makes error only if $d(\theta_i, \hat{\theta}) \geq \frac{d_{\text{min}}}{2} = \psi_n \Rightarrow$

$$\mathbb{P}\left(\hat{i} \neq i \mid \theta_i \text{ true codeword}\right) \leq \mathbb{P}\left(d(\theta_i, \hat{\theta}) \geq \psi_n\right)$$

Therefore

$$\varepsilon_M := \frac{1}{M} \sum_{i=1}^{M} \mathbb{P}\left(\hat{i} \neq i \mid \theta_i \text{ true codeword}\right) \leq \sup_{\theta \in \mathcal{F}} \mathbb{P}\left(d(\theta, \hat{\theta}) \geq \psi_n\right)$$
Probability of decoding error

Minimum distance between codewords is \( d_{\text{min}} = 2\psi_n \Rightarrow \)

Decoder makes error only if \( d(\theta_i, \hat{\theta}) \geq \frac{d_{\text{min}}}{2} = \psi_n \Rightarrow \)

\[
P\left( \hat{i} \neq i \mid \theta_i \text{ true codeword} \right) \leq P\left( d(\theta_i, \hat{\theta}) \geq \psi_n \right)
\]

Therefore

\[
\varepsilon_M := \frac{1}{M} \sum_{i=1}^{M} P\left( \hat{i} \neq i \mid \theta_i \text{ true codeword} \right) \leq \sup_{\theta \in \mathcal{F}} P\left( d(\theta, \hat{\theta}) \geq \psi_n \right)
\]

Plugging into our risk lower bound,

\[
\sup_{\theta \in \mathcal{F}} \mathbb{E} \left[ d(\theta, \hat{\theta}) \right] \geq \psi_n \sup_{\theta \in \mathcal{F}} P\left( d(\theta, \hat{\theta}) \geq \psi_n \right) \geq \psi_n \varepsilon_M
\]
Risk Lower Bound

\[
\sup_{\theta \in \mathcal{F}} \mathbb{E} \left[ d(\theta, \hat{\theta}) \right] \geq \psi_n \varepsilon_M
\]

\( \varepsilon_M \) is average error probability of codebook with \( d_{\text{min}} = 2\psi_n \)

---

Ibragimov and Khasminskii, *Estimation of infinite dimensional parameter in Gaussian white noise*, 1977
Risk Lower Bound

\[
\begin{align*}
\sup_{\theta \in \mathcal{F}} \mathbb{E} \left[ d(\theta, \hat{\theta}) \right] & \geq \psi_n \varepsilon_M \\
\varepsilon_M & \text{ is average error probability of codebook with } d_{\min} = 2\psi_n
\end{align*}
\]

Fano’s inequality is a standard way to lower bound \( \varepsilon_M \):

\[
\varepsilon_M \geq 1 - \frac{\log 2 + \frac{1}{M} \sum_{i=1}^{M} D(P_{Y|\theta_i} \| \overline{P}_Y)}{\log M}, \quad \text{where } \overline{P}_Y = \frac{1}{M} \sum_{i=1}^{M} P_{Y|\theta_i}
\]

If we show that \( \frac{1}{M} \sum_{i=1}^{M} D(P_{Y|\theta_i} \| \overline{P}_Y) \leq \alpha \log M \), then

\[
\varepsilon_M \geq 1 - \alpha - \frac{1}{\log M} > 0,
\]

Ibragimov and Khasminskii, *Estimation of infinite dimensional parameter in Gaussian white noise*, 1977
Improving on Fano

Generalized versions of Fano: [Birgé ’05], [Sason-Verdú ’18]

Other converse techniques:

Sphere-packing bound: [Shannon-Gallager-Berlekamp ’67]

Based on information spectrum: [Wolfowitz ’68], [Verdú-Han ’94]

Based on general $f$-divergences: [Guntuboyina ’11]

Based on binary hypothesis testing:
[Hayashi, Nagaoka ’03]
[Polyanskiy, Poor, Verdú ’10] (“Meta-converse”)  
[Vazquez-Vilar, Tauste Campo, Guillén i Fàbregas, Martinez ’16]
Improving on Fano

Generalized versions of Fano: [Birgé ’05], [Sason-Verdú ’18]

Other converse techniques:

Sphere-packing bound: [Shannon-Gallager-Berlekamp ’67]

Based on information spectrum: [Wolfowitz ’68], [Verdú-Han ’94]

Based on general $f$-divergences: [Guntuboyina ’11]

Based on binary hypothesis testing:

[Hayashi, Nagaoka ’03]

[Polyanskiy, Poor, Verdú ’10] (“Meta-converse”)

[Vazquez-Vilar, Tauste Campo, Guillén i Fàbregas, Martinez ’16]
What we want is a lower bound for $\varepsilon_M$ that:

- is computable for wide range of statistical problems
- with existing packing sets
- shows $\varepsilon_M \to 1$ as $M$ grows (strong converse)
Obtaining a tighter lower bound

- Channel $P_{Y|\theta}$
- Codebook $\{\theta_1, \ldots, \theta_M\}$ (equally likely codewords)
- Average error probability $\varepsilon_M$

A channel decoder defines a hypothesis test to distinguish between:

$$H_0 : (\theta, Y) \sim Q = P_\theta Q_Y$$
$$H_1 : (\theta, Y) \sim P = P_\theta P_{Y|\theta}$$

Does the data look like it came from the true generating model?
Obtaining a tighter lower bound

- Channel $P_{Y|\theta}$
- Codebook $\{\theta_1, \ldots, \theta_M\}$ (equally likely codewords)
- Average error probability $\varepsilon_M$

A channel decoder defines a hypothesis test to distinguish between:

\[
H_0 : (\theta, Y) \sim Q = P_{\theta} Q_Y \\
H_1 : (\theta, Y) \sim P = P_{\theta} P_{Y|\theta}
\]

Does the data look like it came from the true generating model?

For the channel decoder based test [Polyanskiy, Poor, Verdú ’10]:

\[
Q[T = 1] = \frac{1}{M}, \quad P[T = 0] = \varepsilon_M
\]
Obtaining a tighter lower bound

- Channel $P_{Y|\theta}$
- Codebook $\{\theta_1, \ldots, \theta_M\}$ (equally likely codewords)
- Average error probability $\varepsilon_M$

A channel decoder defines a hypothesis test to distinguish between:

$$H_0 : (\theta, Y) \sim Q = P_\theta Q_Y$$
$$H_1 : (\theta, Y) \sim P = P_\theta P_{Y|\theta}$$

Does the data look like it came from the true generating model?

For the channel decoder based test [Polyanskiy, Poor, Verdú ’10]:

$$Q[T = 1] = \frac{1}{M}, \quad P[T = 0] = \varepsilon_M$$

For any randomized hypothesis test $T$ and $\gamma > 0$, we have

$$P[T = 1] - \gamma Q[T = 1] \leq P \left[ \frac{dP}{dQ} > \gamma \right].$$
Hence, in our case, for any $\gamma > 0$

$$\frac{1}{M} \geq \frac{1}{\gamma} \left( 1 - \varepsilon_M - \mathbb{P}_{\theta_Y} \left[ \frac{dP_Y|\theta}{dQ_Y} > \gamma \right] \right)$$

- Can bound $\mathbb{P}_{\theta_Y} \left[ \frac{dP_Y|\theta}{dQ_Y} > \gamma \right]$ in terms of Rényi divergences using Markov inequality type argument

- Can optimize over $\gamma$ to deduce . . .
Theorem

For any $\lambda > 0$, and any distribution $Q_Y$ over $\mathcal{Y}$ (satisfying mild absolute continuity condition),

$$\varepsilon_M \geq 1 - \frac{(1 + \lambda)}{(\lambda M)^{1+\lambda}} \left[ \sum_{i=1}^{M} \frac{1}{M} \exp \left( \lambda D_{1+\lambda}(P_{Y|\theta_{i}} \| Q_Y) \right) \right]^{\frac{1}{1+\lambda}}.$$

Here $D_{1+\lambda}(P_{Y|\theta_{i}} \| Q_Y)$ is the Rényi divergence of order $(1 + \lambda)$:

$$D_{1+\lambda}(P_{Y|\theta_{i}} \| Q_Y) := \frac{1}{\lambda} \log \left( \int_{\mathcal{Y}} \left( \frac{dP_{Y|\theta_{i}}}{dQ_Y} \right)^{1+\lambda} dQ_Y \right).$$
Theorem

For any $\lambda > 0$, and any distribution $Q_Y$ over $\mathcal{Y}$ (satisfying mild absolute continuity condition),

$$\varepsilon_M \geq 1 - \frac{(1 + \lambda)}{(\lambda M)^{\frac{\lambda}{1+\lambda}}} \left[ \sum_{i=1}^{M} \frac{1}{M} \exp \left( \lambda D_{1+\lambda}(P_{Y|\theta_i} \parallel Q_Y) \right) \right]^{\frac{1}{1+\lambda}}.$$ 

Here $D_{1+\lambda}(P_{Y|\theta_i} \parallel Q_Y)$ is the Rényi divergence of order $(1 + \lambda)$:

$$D_{1+\lambda}(P_{Y|\theta_i} \parallel Q_Y) := \frac{1}{\lambda} \log \left( \int_{\mathcal{Y}} \left( \frac{dP_{Y|\theta_i}}{dQ_Y} \right)^{1+\lambda} dQ_Y \right).$$

- Pick a good $Q_Y$, compute lower bound for $\varepsilon_M$ via upper bound for Rényi divergence, e.g., [Sason-Verdú ’16]

- Have free choice of $\lambda$, often $\lambda = 1$ works well enough
Improved risk lower bounds

\[ \sup_{\theta \in \mathcal{F}} \mathbb{E} \left[ d(\theta, \hat{\theta}) \right] \geq \psi_n \varepsilon_M \]

In paper we use the result to study three illustrative examples:

1. Compressed sensing
2. Density estimation problem
3. Active learning of a binary classifier... see paper.

In each case, get improved bounds with \( \varepsilon_M \to 1 \) (strong converse), essentially for free.
Application: Compressed Sensing

\[\mathbf{y} = \mathbf{A} \theta + \mathbf{w}, \quad \mathbf{w} \sim \mathcal{N}(0,\sigma^2\mathbf{I})\]

\(\mathcal{F}_k\): unit norm vectors \(\theta\) in \(\mathbb{R}^n\) with at most \(k\) non-zeros

Want to lower bound

\[M^*(\mathbf{A}) := \inf_{\hat{\theta}} \sup_{\theta \in \mathcal{F}_k} \mathbb{E} \left[ \frac{1}{n} \| \hat{\theta}(\mathbf{y}) - \theta \|^2 \right],\]
Packing set (see [Candès-Davenport ’13])

Packing set of vectors \(\{\theta_1, \ldots, \theta_M\} \in \mathbb{R}^n\) with:

- \(\|\theta_i\|^2 = 1\) for all \(i\)
- \(\|\theta_i - \theta_j\|^2 \geq \frac{1}{2}\) for \(i \neq j\)
- \(\|\frac{1}{M} \sum_{i=1}^{M} \theta_i \theta_i^T - \frac{1}{n} I\|_{\text{op}} \leq \frac{\beta}{n}\) for some small \(\beta > 0\)
Packing set (see [Candès-Davenport '13])

Packing set of vectors \( \{\theta_1, \ldots, \theta_M\} \in \mathbb{R}^n \) with:

- \( \|\theta_i\|^2 = 1 \) for all \( i \)
- \( \|\theta_i - \theta_j\|^2 \geq \frac{1}{2} \) for \( i \neq j \)
- \( \|\frac{1}{M} \sum_{i=1}^{M} \theta_i \theta_i^T - \frac{1}{n} \mathbf{I}\|_{\text{op}} \leq \frac{\beta}{n} \) for some small \( \beta > 0 \)
- Size of packing set \( M = \left( \frac{n}{k} \right)^{k/4} = \exp \left( \frac{k}{4} \log \left( \frac{n}{k} \right) \right) \)
Packing set (see [Candès-Davenport ’13])

Packing set of vectors \(\{\theta_1, \ldots, \theta_M\} \in \mathbb{R}^n\) with:

- \(\|\theta_i\|^2 = 1\) for all \(i\)
- \(\|\theta_i - \theta_j\|^2 \geq \frac{1}{2}\) for \(i \neq j\)
- \(\frac{1}{M} \sum_{i=1}^{M} \theta_i \theta_i^T - \frac{1}{n} I\)\(\|_{op} \leq \frac{\beta}{n}\) for some small \(\beta > 0\)
- Size of packing set \(M = (\frac{n}{k})^{k/4} = \exp\left(\frac{k}{4} \log\left(\frac{n}{k}\right)\right)\)
Computing the Renyi Divergence

\[
\varepsilon_M \geq 1 - \frac{(1 + \lambda)}{(\lambda M)^{\frac{\lambda}{1+\lambda}}} \left[ \sum_{i=1}^{M} \frac{1}{M} \exp \left( \lambda D_{1+\lambda}(P_{Y|\theta_i} \parallel Q_Y) \right) \right]^{\frac{1}{1+\lambda}}
\]

Since \( y = A \theta + w, \) \( P_{Y|\theta_i} \sim \mathcal{N}(A\theta_i, \sigma^2 I) \)
Choose \( Q_Y \sim \mathcal{N}(0, \sigma^2 I) \)
Computing the Renyi Divergence

\[ \varepsilon_M \geq 1 - \frac{(1 + \lambda)}{(\lambda M)^{\frac{\lambda}{1+\lambda}}} \left[ \sum_{i=1}^{M} \frac{1}{M} \exp \left( \lambda D_{1+\lambda}(P_{Y|\theta_i} || Q_Y) \right) \right]^{\frac{1}{1+\lambda}} \]

Since \( y = A \theta + w, \quad P_{Y|\theta_i} \sim \mathcal{N}(A\theta_i, \sigma^2 I) \)

Choose \( Q_Y \sim \mathcal{N}(0, \sigma^2 I) \)

Then

\[ D_{1+\lambda}(P_{Y|\theta_i} || Q_Y) = \frac{(1 + \lambda)}{2\sigma^2} \| A\theta_i \|^2 \]
Computing the Renyi Divergence

\[ \varepsilon_M \geq 1 - \frac{(1 + \lambda)}{(\lambda M)^\frac{1}{1+\lambda}} \left[ \sum_{i=1}^{M} \frac{1}{M} \exp \left( \lambda D_{1+\lambda}(P_{Y|\theta_i} \parallel Q_Y) \right) \right]^{\frac{1}{1+\lambda}} \]

Since \( y = A \theta + w \), \( P_{Y|\theta_i} \sim \mathcal{N}(A\theta_i, \sigma^2 I) \)
Choose \( Q_Y \sim \mathcal{N}(0, \sigma^2 I) \)

Then

\[ D_{1+\lambda}(P_{Y|\theta_i} \parallel Q_Y) = \frac{(1 + \lambda)}{2\sigma^2} \|A\theta_i\|^2 \]

We use a subset \( \mathcal{P} \) of the Candès-Davenport packing set with \( M' = \frac{M}{\log M} \) elements such that

\[ \max_{\theta_i \in \mathcal{P}} \|A\theta_i\|^2 \leq \left(\frac{\|A\|^2_F}{n}\right) (1 + \delta) \quad \text{for some small } \delta > 0 \]
Proposition:
For any $\lambda > 0$, $\Delta \in (0, 1)$, and $M = (n/k)^{k/4}$, we have

$$\varepsilon_M \geq 1 - (1 + \lambda) \left( \frac{(\log M) M^{-\Delta}}{\lambda} \right)^{\lambda/(1+\lambda)},$$
Proposition:

For any $\lambda > 0$, $\Delta \in (0,1)$, and $M = (n/k)^{k/4}$, we have

$$\varepsilon_M \geq 1 - (1 + \lambda) \left( \frac{(\log M)M^{-\Delta}}{\lambda} \right)^{\lambda/(1+\lambda)},$$

and

$$M^*(A) = \inf_{\hat{\theta}} \sup_{\theta \in \mathcal{F}_k} \mathbb{E} \left[ \frac{1}{n} \|\hat{\theta}(y) - \theta\|^2 \right] \geq \frac{\sigma^2}{4\|A\|^2_F} \left( \frac{k}{4} \log \frac{n}{k} - 1 \right) \frac{(1 - \Delta)}{(1 + \lambda)} \varepsilon_M,$$
For large $n$ we have

$$M^*(A) \geq \frac{\sigma^2}{4\|A\|_F^2} \left( \frac{k}{4} \log \frac{n}{k} \right) (1 - o(1)).$$

改善因子约为8，超过Fano论证中的Candès-Davenport，该论证给出了

$$M^*(A) \geq \frac{\sigma^2}{32\|A\|_F^2 (1 + \beta)} \left( \frac{k}{4} \log \frac{n}{k} - 2 \right).$$

MSE的相同数量级可以实现与$A$满足RIP的现象，改善超越常数因子是不可能的。
Consider $\mathcal{F}$, set of probability densities $\theta$ on $[0, 1]$ such that

$$a_0 \leq \theta \leq a_1 \quad \text{and} \quad |\theta''(x)| \leq a_2$$

We are given $(Y_1, \ldots, Y_n)$ generated IID from $\theta$.

Wish to estimate density with $\hat{\theta}_n = \hat{\theta}_n(Y_1, \ldots, Y_n)$.

Measure performance by squared Hellinger distance

$$d(\theta, \hat{\theta}_n) = \int_0^1 \left( \sqrt{\theta(x)} - \sqrt{\hat{\theta}_n(x)} \right)^2 \, dx.$$ 

Wish to obtain lower bound on minimax risk $\inf_{\hat{\theta}_n} \sup_{\theta \in \mathcal{F}} \mathbb{E}d(\theta, \hat{\theta}_n)$
Packing set (see Yu ’97)

Packing set consists of densities that are small perturbations of uniform density on \([0, 1]\)

- Fix a smooth, bounded \(g(x)\) with
  \[
  \int_0^1 g(x)dx = 0 \quad \text{and} \quad \int_0^1 (g(x))^2 dx = a.
  \]

- Partition \([0, 1]\) into \(m\) subintervals of length \(1/m\)
- Perturb uniform density in each subinterval by small amount proportional to rescaled version of \(g\)
- That is, for some \(c\) define
  \[
  g_j(x) = \frac{c}{m^2} g(mx-j) \mathbb{1} \left( \frac{j}{m} \leq x < \frac{j+1}{m} \right), \quad \text{for } j = 0, \ldots, m - 1.
  \]
Packing set (contd.)

- Hypercube class of $2^m$ densities

\[
\left\{ f_\tau(y) = 1 + \sum_{j=0}^{m-1} \tau_j g_j(y) : \tau = (\tau_1, \ldots, \tau_m) \in \{-1,1\}^m \right\}
\]

(In subinterval $j$, perturb uniform by either $g_j$ or $-g_j$)

- Bandwidth parameter $m$ chosen later to optimize lower bound
Packing set (contd.)

- Hypercube class of $2^m$ densities

$$\left\{ f_{\tau}(y) = 1 + \sum_{j=0}^{m-1} \tau_j g_j(y) : \tau = (\tau_1, \ldots, \tau_m) \in \{\pm 1\}^m \right\}$$

(In subinterval $j$, perturb uniform by either $g_j$ or $-g_j$)

- Bandwidth parameter $m$ chosen later to optimize lower bound

Pick packing set corresponding to well-separated sequences in $\{-1, 1\}^m$ (guaranteed by Gilbert-Varshamov)

- $\mathcal{A} \subseteq \{-1, 1\}^m$ whose elements have pairwise Hamming distance $\geq m/3$

- Size of $\mathcal{A} \geq \exp(c_0m)$, where $c_0 \simeq 0.082$

- Resulting packing set $\{f_{\tau} : \tau \in \mathcal{A}\}$ has minimum squared Hellinger distance $d_{\text{min}} = ac^2/(3m^4)$ (see Bin Yu)
Using main theorem

For $Q_Y$ uniform and $\lambda = 1$ in main theorem, Rényi term is

$$\left[ \sum_{\tau \in \mathcal{A}} \frac{1}{M} \int_{[0,1]^n} f^n_{\tau}(y)^2 \, dy \right]^{\frac{1}{2}} \leq \exp \left( \frac{c^2 an}{2m^4} \right).$$
Proposition:
With $m = n^{1/5}/\nu$ for any positive constant $\nu < (c_0/(c^2a))^{1/5}$, the minimax risk satisfies

$$\inf_{\hat{\theta}_n} \sup_{\theta \in \mathcal{F}} \mathbb{E} d(\hat{\theta}_n, \theta) \geq \frac{c^2 a \nu^4}{6} n^{-4/5} \varepsilon_M,$$

where

$$\varepsilon_M \geq 1 - 2 \exp \left( \frac{-n^{1/5}}{2\nu} (c_0 - \nu^5 c^2a) \right).$$

- Bin Yu method uses same packing set + Fano, but gives $\varepsilon_M$ bounded away from zero, not converging to 1
Summary

Lower bounds on minimax risk: packing set + lower bound on $\varepsilon_M$

- Computable via bounding Rényi divergence, gives strong converse
- Other example in paper: active learning of binary classifier
- Improvements over main theorem possible (Baraud arxiv:1807.05410)

Further work:

- Can this method give improved minimax rates, rather than just improved constants?
- Extend results to work with global metric entropy features [Yang-Barron '99], [Guntuboyina '11]

Paper in *Electronic Journal of Statistics* (OA), 2018
doi: 10.1214/18-EJS1419