Electronic Journal of Statistics
ISSN: 1935-7524
arXiv: 1706.04410

A strong converse bound for multiple hypothesis testing, with applications to high-dimensional estimation

Ramji Venkataramanan* **and Oliver Johnson†**

Abstract: In statistical inference problems, we wish to obtain lower bounds on the minimax risk, that is to bound the performance of any possible estimator. A standard technique to do this involves the use of Fano's inequality. However, recent work in an information-theoretic setting has shown that an argument based on binary hypothesis testing gives tighter converse results (error lower bounds) than Fano for channel coding problems. We adapt this technique to the statistical setting, and argue that Fano's inequality can always be replaced by this approach to obtain tighter lower bounds that can be easily computed and are asymptotically sharp. We illustrate our technique in three applications: density estimation, active learning of a binary classifier, and compressed sensing, obtaining tighter risk lower bounds in each case.

MSC 2010 subject classifications: 62G05, 62B10, 62G07.
Keywords and phrases: minimax lower bounds, Fano's inequality, compressed sensing, density estimation, active learning.

1. Introduction

When solving an inference problem, we would like to know if the algorithm we use is close to optimal. In statistical language we seek to give a lower bound on the performance of any estimator over a class of problems (often called the minimax risk over the class). In the language of information theory, we speak of converse results, which give performance bounds for all communication schemes over a noisy channel.

In the statistics literature, one standard approach to proving converse results is via Fano's inequality (see [11, Theorem 2.11.1]). However, recent information-theoretic literature has shown how to obtain sharper converse bounds. The resulting improvements can be significant at finite sample size, and give bounds that are close to optimal, as illustrated in the work of Polyanskiy, Poor and Verdú [23]. The present paper shows how the method of [23], although developed for channel coding problems, gives stronger risk lower bounds for high-dimensional estimation problems, compared to the standard Fano approach.

*Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB3 0DZ, UK. Email: ramji.v@eng.cam.ac.uk

†School of Mathematics, University of Bristol, University Walk, Bristol, BS8 1TW, UK. Email: maotj@bristol.ac.uk

We first describe the general set-up, following the treatment and notation of [31, Chapter 2]. Consider an inference problem (possibly non-parametric) where we wish to estimate some $\theta \in \mathcal{F}$ from samples $\mathbf{Y} = (Y_1, \ldots, Y_n)$ generated according to a distribution $P_\theta(\mathbf{Y})$. For example, in Section 3 we consider θ to be a probability density chosen from a pre-specified class, and in Section 5 we consider θ to be a k-sparse vector in $\mathbb{R}^n$. Let $\widehat{\theta} := \widehat{\theta}(\mathbf{Y})$ be any estimator of θ and let $d(\theta, \widehat{\theta})$ represent the loss. We assume that d is a distance, although (as in [31]) our results hold when d is a semi-distance; that is, when $d(\theta, \theta') = 0$ need not imply that $\theta = \theta'$. We obtain lower bounds on the minimax risk

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathcal{F}} \mathbb{E}\left[w(d(\theta, \widehat{\theta}))\right], \tag{1}$$

where w is any monotonically increasing function with $w(0) = 0$. For example, we may consider $w(u) = u^p$ for any $p > 0$ or $w(u) = \mathbb{I}(u \geq c)$ for some $c > 0$.

A standard approach for obtaining a lower bound on (1) is as follows. First, a set $\{\theta_1, \ldots, \theta_M\} \subseteq \mathcal{F}$ is chosen, with a lower bound on the pairwise distance between any two of its elements, where the distance is measured using the loss function $d(\cdot, \cdot)$. Then, any estimator $\widehat{\theta}$ defines an M-ary hypothesis test that detects one of $\{\theta_1, \ldots, \theta_M\}$ based on the data $\mathbf{Y}$. Next, the key step is to obtain a lower bound for the error probability associated with this hypothesis test. For a well-constructed set, Fano's inequality often shows that this average error probability is bounded away from 0 as $n \to \infty$. In this paper, we present a technique that often shows that it approaches 1 as $n \to \infty$. In information theory parlance (see for example [11, P.207]), we prove a "strong converse" result in contrast to the "weak converse" provided by Fano's inequality.

We now explain the details, following the framework in [31] (see also [13], [20]). For any positive constants A and ψ_n, using Markov's inequality we have

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

which implies

$$\sup_{\theta \in \mathcal{F}} \mathbb{E}\left[w\left(\frac{1}{\psi_n} d(\theta, \widehat{\theta})\right)\right] \geq w(A)\left(\sup_{\theta \in \mathcal{F}} \mathbb{P}\left(d(\theta, \widehat{\theta}) \geq A\psi_n\right)\right). \tag{2}$$

When applying (2), we typically choose ψ_n as a decreasing function of n to give the desired convergence rate, and A as a constant that can be used to optimize the lower bound. The goal then is to control the bracketed term on the RHS of (2) to obtain a lower bound on the minimax risk. We use the following definition.

Definition 1.1. *A collection $\mathcal{P}_{M,d_{\min}} = \{\theta_1, \ldots, \theta_M\} \subseteq \mathcal{F}$ is called a packing set of size M and minimum distance $d_{\min}$ if*

$$d(\theta_i, \theta_j) \geq d_{\min}, \qquad \textit{for all } i \neq j.$$

In general, the packing set is not explicitly constructed, but its existence is guaranteed via combinatorial arguments. In Remarks 3.1 and 4.1 below, existence of packing sets is guaranteed by applying the Gilbert–Varshamov bound. In Remark 5.1, the existence of a packing set is guaranteed via the probabilistic method. We emphasize that we use these existing packing set constructions: our contribution is to provide tighter lower bounds than can be obtained using Fano's inequality. It is possible that the resulting risk lower bounds could be improved by a further constant factor, by varying the packing set construction.

In statistical language, we think of the packing set $\mathcal{P}_{M,d_{\min}}$ as multiple hypotheses to be distinguished on the basis of data. An alternative information-theoretic interpretation is to think of $\mathcal{P}_{M,d_{\min}}$ as a codebook, that is a collection of M codewords, one of which is transmitted over a noisy communication channel. Given a packing set $\mathcal{P}_{M,d_{\min}}$, any estimator $\widehat{\theta}$ provides a way to distinguish between multiple hypotheses (act as a channel decoder) as follows: given $\widehat{\theta}$, we choose $\widehat{i} = \arg\min_j d(\widehat{\theta}, \theta_j)$, i.e. the index of the closest value in the packing set. In coding theory, this is called the minimum distance decoder.

If θ_i is the true value, a simple triangle inequality argument shows that $\left\{\widehat{i} \neq i\right\} \Rightarrow \left\{d(\theta_i, \widehat{\theta}) \geq d_{\min}/2\right\}$. Taking $d_{\min} = 2A\psi_n$, we can bound the bracketed term on the RHS of (2) by the average error probability ϵ_M of the optimal decoder $i^* = i^*(\mathbf{Y})$, since

$$\begin{aligned}
\sup_{\theta\in\mathcal{F}} \mathbb{P}\left(d(\theta,\widehat{\theta}) \geq A\psi_n\right) &\geq \max_{i\in\{1,\ldots,M\}} \mathbb{P}\left(d(\theta_i,\widehat{\theta}) \geq A\psi_n\right) \\
&\geq \max_{i\in\{1,\ldots,M\}} \mathbb{P}(\theta_{\widehat{i}} \neq \theta_i) \\
&\geq \max_{i\in\{1,\ldots,M\}} \mathbb{P}(\theta_{i^*} \neq \theta_i) \\
&\geq \frac{1}{M}\sum_{i=1}^{M} \mathbb{P}(\theta_{i^*} \neq \theta_i) =: \epsilon_M. \qquad (3)
\end{aligned}$$

This calculation and argument are standard in the literature (see for example [31, Eq. (2.9)], [20, Corollary 2.19]). By substituting (3) in (2), we deduce

$$\inf_{\widehat{\theta}} \sup_{\theta\in\mathcal{F}} \mathbb{E}\left[w\left(\frac{1}{\psi_n} d(\theta,\widehat{\theta})\right)\right] \geq w(A)\,\epsilon_M. \qquad (4)$$

Our main focus is to obtain a sharp and easily computable bound for ϵ_M. A standard technique, dating back to Ibragimov and Khasminskii [17], is to bound ϵ_M using Fano's inequality, which gives the bound [31, Lemma 2.10]

$$\epsilon_M \geq 1 - \frac{\log 2 + \frac{1}{M}\sum_{i=1}^{M} D(P_{\theta_i}\|\overline{P})}{\log M}, \qquad (5)$$

where $\overline{P} := \frac{1}{M}\sum_{i=1}^{M} P_{\theta_i}$, and $D(P\|Q)$ is the Kullback–Leibler (KL) divergence. To apply (5), one typically obtains a bound of the form

$$\frac{1}{M}\sum_{i=1}^{M} D(P_{\theta_i}\|\overline{P}) \leq \alpha \log M,$$

for some constant $\alpha \in (0,1)$ (see [31, Section 2.7.1]). Then (5) implies that $\epsilon_M \geq 1-\alpha-\frac{\log 2}{\log M}$, which converges to $(1-\alpha)$ for large sample sizes n (assuming $\log M \to \infty$ as $n \to \infty$), meaning that we deduce a weak converse, and (3) gives a lower bound on the risk (via (2)).

The remainder of the paper is organized as follows. In Section 2, we derive a lower bound on ϵ_M (Theorem 1) that strengthens Fano's inequality. In Section 2.1, we discuss related prior work. We then apply Theorem 1 to three high-dimensional estimation problems, in each case showing the average error probability $\epsilon_M \to 1$ as $n \to \infty$ (strong converse). In each case, our method replaces the Fano-based part of the argument which gives a weak converse. In Section 3, we give a strong converse for a density estimation problem studied by Yu [35]. In Section 4, we obtain strengthened risk lower bounds for active learning of a binary classifier, following Castro and Nowak [10]. In Section 5, we use Theorem 1 to improve (by a factor of nearly 8) lower bounds of Candès and Davenport [8] for the minimax mean-squared error in compressed sensing.

2. Lower bound on the Average Error Probability

We bound the average error probability ϵ_M in (3) using a different *binary* hypothesis testing problem. Adopting the formalism of [23], we consider a random variable S representing a message chosen uniformly at random from $\{1,\dots,M\}$. The message S is acted on by the simple encoder that generates codeword $\theta = \theta_S$, giving an induced distribution π_θ uniform over the set $\{\theta_1,\dots,\theta_M\}$.

We think of $\mathbf{Y} = (Y_1,\dots,Y_n)$ as the output of a channel with input θ. Using arguments from [23, 32], we bound the desired average error probability of the optimal decoder (3) in terms of the Type I error probability of the following binary hypothesis testing problem:

$$H_0 : (\theta, \mathbf{Y}) \sim Q_{\theta\mathbf{Y}} := \pi_\theta Q_{\mathbf{Y}} \tag{6}$$

$$H_1 : (\theta, \mathbf{Y}) \sim P_{\theta\mathbf{Y}} := \pi_\theta P_{\mathbf{Y}|\theta}, \tag{7}$$

for some probability distribution $Q_{\mathbf{Y}}$ that does not depend on θ. In other words, we wish to determine whether θ and $\mathbf{Y}$ are independent, or are generated by the true underlying channel model. We assume that the measure $Q_{\mathbf{Y}}$ dominates $P_{\mathbf{Y}|\theta_i}$ for $1 \leq i \leq M$, and hence the Radon-Nikodym derivative $\frac{dP_{\mathbf{Y}|\theta_i}}{dQ_{\mathbf{Y}}}$ exists.

The space of $\mathbf{Y}$ is denoted by $\mathcal{Y}$. Throughout the paper, we use boldface notation to denote vectors of length n.

Theorem 1. *Let ϵ_M denote the average error probability of any decoder over channel $P_{\mathbf{Y}|\theta}$, for a channel code with input distribution π_θ uniform over the M codewords $\{\theta_1,\dots,\theta_M\}$. For any $\lambda > 0$, and any distribution $Q_{\mathbf{Y}}$ over $\mathcal{Y}$ such that $P_{\mathbf{Y}|\theta_i}$ is absolutely continuous with respect to $Q_{\mathbf{Y}}$ for $1 \leq i \leq M$,*

$$\epsilon_M \geq 1 - \frac{(1+\lambda)}{(\lambda M)^{\frac{\lambda}{1+\lambda}}} \left[\sum_{i=1}^{M} \frac{1}{M} \exp\left(\lambda D_{1+\lambda}(P_{\mathbf{Y}|\theta_i} \| Q_{\mathbf{Y}})\right)\right]^{\frac{1}{1+\lambda}}. \tag{8}$$

Here $D_{1+\lambda}(P_{\mathbf{Y}|\theta_i}\|Q_{\mathbf{Y}})$ is the Rényi divergence of order $(1+\lambda)$ defined as

$$D_{1+\lambda}(P_{\mathbf{Y}|\theta_i}\|Q_{\mathbf{Y}}) := \frac{1}{\lambda}\log\left(\int_{\mathcal{Y}}\left(\frac{dP_{\mathbf{Y}|\theta_i}}{dQ_{\mathbf{Y}}}\right)^{1+\lambda} dQ_{\mathbf{Y}}\right). \tag{9}$$

The proof uses the following lemma, itself proved in Appendix A.

Lemma 2.1. *With the assumptions and notation of Theorem 1 we have for any $\gamma > 0$*

$$\frac{1}{M} \geq \frac{1}{\gamma}\left(1 - \epsilon_M - P_{\theta\mathbf{Y}}\left[\frac{dP_{\mathbf{Y}|\theta}}{dQ_{\mathbf{Y}}} > \gamma\right]\right). \tag{10}$$

Proof of Theorem 1. Writing $\mathbb{I}(\cdot)$ for the indicator function, the probability in (10) satisfies

$$\begin{aligned} P_{\theta\mathbf{Y}}\left[\frac{dP_{\mathbf{Y}|\theta}}{dQ_{\mathbf{Y}}} > \gamma\right] &= \int_{\mathcal{Y}}\sum_{i=1}^{M}\frac{1}{M}\,\mathbb{I}(\theta=\theta_i)\,\frac{dP_{\mathbf{Y}|\theta}}{dQ_{\mathbf{Y}}}(\mathbf{y})\,\mathbb{I}\left[\frac{dP_{\mathbf{Y}|\theta}}{dQ_{\mathbf{Y}}}(\mathbf{y}) > \gamma\right] dQ_{\mathbf{Y}}(\mathbf{y}) \\ &= \int_{\mathcal{Y}}\sum_{i=1}^{M}\frac{1}{M}\,\frac{dP_{\mathbf{Y}|\theta_i}}{dQ_{\mathbf{Y}}}(\mathbf{y})\,\mathbb{I}\left[\frac{dP_{\mathbf{Y}|\theta_i}}{dQ_{\mathbf{Y}}}(\mathbf{y}) > \gamma\right] dQ_{\mathbf{Y}}(\mathbf{y}) \\ &\leq \int_{\mathcal{Y}}\sum_{i=1}^{M}\frac{1}{M}\,\frac{dP_{\mathbf{Y}|\theta_i}}{dQ_{\mathbf{Y}}}(\mathbf{y})\left(\frac{1}{\gamma}\frac{dP_{\mathbf{Y}|\theta_i}}{dQ_{\mathbf{Y}}}(\mathbf{y})\right)^{\lambda} dQ_{\mathbf{Y}}(\mathbf{y}) \qquad \text{for } \lambda > 0. \end{aligned} \tag{11}$$

Using this bound (11) in Lemma 2.1, we have

$$\frac{1}{M} \geq \sup_{\gamma>0}\left[\frac{1-\epsilon_M}{\gamma} - \frac{1}{\gamma^{1+\lambda}}\sum_{i=1}^{M}\frac{1}{M}\int_{\mathcal{Y}}\left(\frac{dP_{\mathbf{Y}|\theta_i}}{dQ_{\mathbf{Y}}}(\mathbf{y})\right)^{1+\lambda} dQ_{\mathbf{Y}}(\mathbf{y})\right]. \tag{12}$$

Computing the maximum over $\gamma > 0$ and rearranging, we get (8). □

Remark 2.1. *As shown in the active learning example in Sec. 4, one can use upper bounds for the Rényi divergence in (9) to obtain lower bounds for ϵ_M. Such upper bounds can found, for example, in [26, 27].*

Remark 2.2. *In Appendix B, we show how Fano's inequality in (5) can be obtained from the lower bound on Theorem 1. Furthermore, the examples in the next three sections show that Theorem 1 yields strictly better lower bounds than the Fano-based approach.*

Remark 2.3. *If we assume that each $P_{\mathbf{Y}|\theta_i}$ has a density $p_{\theta_i}(\mathbf{y})$ with respect to a common reference measure μ, then the choice of $Q_{\mathbf{Y}}$ that maximizes the lower bound in (8) has the following density with respect to μ [22, 29]:*

$$q^*(\mathbf{y}) = \frac{1}{C}\left(\sum_{i=1}^{M}\frac{1}{M}(p_{\theta_i}(\mathbf{y}))^{1+\lambda}\right)^{\frac{1}{1+\lambda}},$$

where the normalizing constant $C = \int_{\mathcal{Y}} \left(\sum_{i=1}^{M} \frac{1}{M} (p_{\theta_i}(\mathbf{y}))^{1+\lambda} \right)^{\frac{1}{1+\lambda}} d\mu(\mathbf{y})$. *However, the bound in* (8) *is generally not computable with this choice of* $Q_{\mathbf{Y}}$. *As we will see in the following sections, the structure of the problem often suggests a natural choice for* $Q_{\mathbf{Y}}$ *that yields a computable lower bound.*

2.1. Related work

In [31, Proposition 2.2], Tsybakov gives a result similar to Lemma 2.1. This result can then be used to obtain a lower bound on ϵ_M using the average pairwise χ^2-distance between q and the elements of the packing set [31, Theorem 2.6]. This bound is similar to the one obtained by using $\lambda = 1$ in Theorem 1. In this paper, we show that via two examples (active learning and compressed sensing) that Theorem 1 can be applied with a general $\lambda > 0$ to obtain stronger non-asymptotic bounds. Furthermore, as $n \to \infty$, Theorem 1 gives a strong converse ($\epsilon_M \to 1$), unlike Fano's inequality.

Birgé [5] gives stronger, but less transparent, bounds than Fano's inequality using Fano-type arguments; again, ϵ_M is bounded in terms of an average of Kullback–Leibler divergences, but these are used as the argument for a function, rather than directly substituted. Sason and Verdú [28, Section 3] recently derived a generalized Fano's inequality in terms of the Arimoto-Rényi conditional entropy. They also obtained upper bounds on ϵ_M in terms of the pairwise Rényi divergences [28, Section 4].

Note that an alternative approach to hypothesis testing bounds, avoiding the use of Fano's inequality, is given by Assouad [2]. Indeed, [35] makes a detailed comparison between Fano-based bounds and those coming from Assouad's Lemma [2], finding little practical difference. Indeed [35] quotes Birgé [4, p. 279]: "[Fano] is in a sense more general because it applies in more general situations. It could also replace Assouad's Lemma in almost any practical case ...".

Using Fano's inequality, Yang and Barron [33] obtained order-optimal minimax risk lower bounds that depend only on global metric entropy features of the underlying function class, without explicitly constructing a packing set. The required metric entropy features (bounds on the packing number and covering number) are available from results in approximation theory for many function classes of interest. Guntuboyina [14] obtained a lower bound on the average error probability in terms of general f-divergences, and also generalized the metric entropy results of Yang and Barron [33] to certain f-divergences such as the χ^2-divergence. An interesting direction for future work would be to obtain a non-asymptotic result analogous to Theorem 1 for the case where only the global metric entropy features are available.

An important historical remark is that Hayashi and Nagaoka [16] first linked channel coding and binary hypothesis testing, with later work [15] by Hayashi clarifying this approach and Nagaoka [21] using similar ideas to derive strong converse results. The recent work by Vazquez-Vilar et al. [32] also provides results characterizing the average error probability of channel coding in terms of the Type I error of a binary hypothesis test. This link with channel coding

has been used in other contexts to prove strong converse results, including [18], which derived strong converse results for the group testing problem.

3. Application to density estimation

For the remainder of this paper, we show how Theorem 1 can be applied to a number of high-dimensional estimation problems. In this section we apply Theorem 1 to the following density estimation problem taken from Yu [35, Example 2, P.431]. Let $\mathcal{F}$ be the class of smooth densities on $[0,1]$ such that for any density $\theta \in \mathcal{F}$, we have

$$\int_0^1 \theta(x)dx = 1, \qquad a_0 \le \theta(x) \le a_1 < \infty, \qquad |\theta''(x)| \le a_2, \quad x \in \mathbb{R},$$

for some positive constants a_0, a_1, a_2. The goal is to estimate the density θ from $\mathbf{Y} = (Y_1, \ldots, Y_n)$, where $\{Y_i\}$ are generated IID from θ. We want to bound from below the risk of any estimator $\widehat{\theta}_n = \widehat{\theta}_n(\mathbf{Y})$, where the loss is measured using squared Hellinger distance, i.e.,

$$d(\theta, \widehat{\theta}_n) = \int_0^1 \left(\sqrt{\theta(x)} - \sqrt{\widehat{\theta}_n(x)}\right)^2 dx. \tag{13}$$

The packing set in [35] is constructed via a hypercube class of densities defined via small perturbations of the uniform density on $[0,1]$. Fix a smooth, bounded function $g(x)$ with

$$\int_0^1 g(x)dx = 0 \quad \text{and} \quad \int_0^1 (g(x))^2\, dx = a. \tag{14}$$

We partition the unit interval $[0,1]$ into m subintervals of length $1/m$, and perturb the uniform density on each subinterval by a small amount, proportional to a version of g rescaled and translated to lie on that subinterval. That is, for some sufficiently small fixed constant c, we can define the functions

$$g_j(x) = \frac{c}{m^2} g(mx - j)\mathbb{I}\left(\frac{j}{m} \le x < \frac{j+1}{m}\right), \quad \text{for } j = 0, \ldots, m-1\ . \tag{15}$$

Considering perturbations of the uniform density by $\pm\{g_j\}$, define the following hypercube class of joint densities indexed by $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_m) \in \{\pm 1\}^m$:

$$\mathcal{M}_m = \left\{ f_{\boldsymbol{\tau}}(y) = 1 + \sum_{j=0}^{m-1} \tau_j g_j(y) \right\}. \tag{16}$$

The bandwidth parameter m will be chosen later as an increasing function of n, to optimize the risk lower bound.

Remark 3.1 (Packing set construction)**.** *[35, Lemma 4] There exists a subset $\mathcal{A} \subseteq \{-1,1\}^m$ of size $M \geq \exp(c_0 m)$, where $c_0 \simeq 0.082$, whose elements have minimum pairwise Hamming distance at least $m/3$. It is then shown in [35] that this results in a packing set of densities*

$$\mathcal{P}_{M,\frac{ac^2}{3m^4}} = \{f_{\boldsymbol{\tau}} : \boldsymbol{\tau} \in \mathcal{A}\} \subseteq \mathcal{M}_m,$$

where $ac^2/(3m^4)$ is a lower bound on the squared Hellinger distance (see (13)) between distinct densities in the packing set. Here a is defined in (14), and c is defined in (15). We use exactly this packing set $\mathcal{P}_{M,ac^2/(3m^4)}$ as the set of M codewords $\{\theta_1, \ldots, \theta_M\}$ in Theorem 1.

We now use Theorem 1 to bound the risk. To do this, we first state an explicit bound (to be proved in Appendix C) on the bracketed term in (8), for $\lambda = 1$.

Lemma 3.1. *Taking $Q_{\mathbf{Y}}$ to be the uniform measure on $[0,1]^n$ and identifying each $\frac{dP_{\mathbf{Y}|\theta_i}}{dQ_{\mathbf{Y}}}$ with a density $f^n_{\boldsymbol{\tau}}(\mathbf{y}) = \prod_{i=1}^n f_{\boldsymbol{\tau}}(y_i)$ for $\tau \in \mathcal{A}$, with $\lambda = 1$, the bracketed term in (8) becomes:*

$$\left[\sum_{i=1}^{M} \frac{1}{M} \int_{\mathcal{Y}} \left(\frac{dP_{\mathbf{Y}|\theta_i}}{dQ_{\mathbf{Y}}}(\mathbf{y})\right)^2 dQ_{\mathbf{Y}}(\mathbf{y})\right]^{\frac{1}{2}} = \left[\sum_{\boldsymbol{\tau}\in\mathcal{A}} \frac{1}{M} \int_{[0,1]^n} f^n_{\boldsymbol{\tau}}(\mathbf{y})^2 d\mathbf{y}\right]^{\frac{1}{2}} \leq \exp\left(\frac{c^2 an}{2m^4}\right).$$

Combining Lemma 3.1 with Theorem 1, we deduce the following lower bound.

Proposition 3.2. *For any positive constant $\nu < (c_0/(c^2 a))^{1/5}$, the risk of any estimator $\widehat{\theta}_n$ satisfies*

$$\sup_{\theta\in\mathcal{F}} \mathbb{E}d(\widehat{\theta}_n, \theta) \geq \frac{c^2 a \nu^4}{6} n^{-4/5} \epsilon_M, \tag{17}$$

where

$$\epsilon_M \geq 1 - 2\exp\left(-\frac{n^{1/5}}{2\nu}\left(c_0 - \nu^5 c^2 a\right)\right). \tag{18}$$

Therefore, for large n we have

$$\sup_{\theta\in\mathcal{F}} \mathbb{E}d(\widehat{\theta}_n, \theta) \geq \frac{c_0^{4/5}(c^2 a)^{1/5}}{6} n^{-4/5}(1 - o(1)). \tag{19}$$

Proof. We apply Theorem 1 by setting the minimum distance $ac^2/(3m^4)$ of the packing set in Remark 3.1 to $2A\psi_n$. Taking $A = 1$, we obtain $\psi_n = c^2 a/(6m^4)$. Taking w to be the identity in (4), we deduce

$$\max_j \ \mathbb{E}d(\hat{\theta}, \theta_j) \geq \psi_n \epsilon_M = \frac{c^2 a}{6m^4} \epsilon_M.$$

Taking $\lambda = 1$ in Theorem 1 and using Lemma 3.1, we bound ϵ_M as

$$\epsilon_M \geq 1 - \frac{2}{\sqrt{M}} \exp\left(\frac{c^2 an}{2m^4}\right) \overset{(a)}{\geq} 1 - 2\exp\left(-\frac{c_0 m}{2}\right) \exp\left(\frac{c^2 an}{2m^4}\right),$$

where (a) is obtained using the fact that the packing set of densities $\mathcal{P}_{M,ac^2/(3m^4)}$ has size $M \geq \exp(c_0 m)$, as described in Remark 3.1 above. We therefore have

$$\max_j \ \mathbb{E}d(\hat{\theta}, \theta_j) \geq \frac{c^2 a}{6m^4}\left[1 - 2\exp\left(-\frac{m}{2}\left(c_0 - \frac{c^2 an}{m^5}\right)\right)\right]. \tag{20}$$

The result (17) follows by taking $m = n^{1/5}/\nu$.

To obtain the asymptotic bound in (19), we take ν to approach $(c_0/(c^2 a))^{1/5}$ as $n \to \infty$, but slowly enough that ensure that the exponent on the RHS of (18) is negative so that ϵ_M tends to 1. For example, we can take $\nu = \left(\frac{c_0}{c^2 a}\right)^{1/5}(1 - n^{-1/\kappa})$ for $\kappa > 25$. □

Remark 3.2. *The paper [35] derives Fano-type bounds in this setting: combining Lemmas 3 and 5 of [35] and taking taking $m = n^{1/5}/\nu$ gives the same bound as (17), but with a looser lower bound on ϵ_M given by*

$$\epsilon_M \geq \left(1 - \frac{1}{c_0}\left(\frac{2c^2 a\nu^5}{(1 - c_g)} + \frac{\log 2}{n^{1/5}}\right)\right). \tag{21}$$

For the bound (21) to be meaningful, we need $\nu < \left(\frac{c_0}{c^2 a}\frac{1-c_g}{2}\right)^{1/5}$, where $c_g = c\sup_x |g(x)|$. The scaling factor c has to be chosen so that $c_g < 1$.

Thus Proposition 3.2 provides a strong converse (error probability tending to 1), whereas the result (21) extracted from [35] gives a weak converse (error probability bounded away from 0). Our bound also offers greater flexibility in choosing ν and removes the need to control c_g.

Remark 3.3. *Theorem 1 can similarly be applied to obtain strong converses for estimating densities belonging to either Hölder or Sobolev classes, strengthening the risk lower bounds described in [31, Sec. 2.6.1]*

4. Application to active learning of a classifier

In this section, we use Theorem 1 to derive strengthened minimax lower bounds for active learning algorithms for a family of classification problems introduced by Castro and Nowak [10] (see also [30]). We use the explicit packing set construction of [10], but modify their notation for consistency.

Consider data of the form $\mathbf{Y} = (\mathbf{U}, \mathbf{V}) = ((U_1, V_1), \ldots, (U_n, V_n))$. Each pair (U_r, V_r) consists of a feature vector $U_r \in \mathbb{R}^d$ (where we assume $d \geq 2$) and a binary label $V_r \in \{0, 1\}$, and is drawn independently from an underlying joint distribution $P_{UV} = P_U P_{V|U}$. The goal of classification is to predict the value of label V, given a future U observation. This is done via G, a measurable subset of

$\mathbb{R}^d$. Given a $U \in \mathbb{R}^d$, the classifier estimates its label as $\hat{V} := \hat{V}(U) := \mathbb{I}(U \in G)$. The risk of a classifier is the probability of classification error, given by

$$R(G) = \mathbb{P}(\hat{V} \neq V) = \mathbb{P}(\mathbb{I}(U \in G) \neq V),$$

where $(U, V) \sim P_{UV}$. It is well-known (see [30]) that, given knowledge of P_{UV}, the Bayes-optimal classifier is

$$G^* = \{u \in \mathbb{R}^d : \eta(u) \geq 1/2\},$$

where the feature conditional probability $\eta(u) = P_{V|U}(1|u)$. As P_{UV} is unknown, our goal is to estimate G^* from data $\mathbf{Y}$. The performance of classifier $\widehat{G}_n$ is measured by *excess risk* (or regret) [10, Eq. (1)]

$$R(\widehat{G}_n) - R(G^*) = \int_{\widehat{G}_n \Delta G^*} |2\eta(u) - 1| \, dP_U(u),$$

where Δ represents the symmetric difference between sets. For the remainder of this section, as in [10], we will assume that P_U is supported on $[0,1]^d$. It is clear that the difficulty of a classification problem will depend on both the shape of the boundary of G^* and the behaviour of $(2\eta(u) - 1)$ for u close to this boundary. We consider the class of joint distribution functions $\mathsf{BF}(\alpha, \kappa, L, c)$, defined formally in [10, Section IV]. For our purposes it suffices to understand this class as a set of distributions P_{UV} such that:

1. The boundary of G^* can be expressed as an α-Hölder smooth function with constant L.
2. The value of $|\eta(u) - 1/2|$ is at least $cD^{\kappa-1}$ for points u at distance D from the boundary, where $\kappa \geq 1$.

Algorithms that attempt to learn the Bayes-optimal classifier G^* from data are categorized as passive or active. *Passive* learning algorithms aim to learn G^* from a pre-specified, possibly random, choice of $(U_1, \ldots, U_n)$ and the corresponding labels $(V_1, \ldots, V_n)$. In contrast, *active* learning algorithms choose each U_r based on previous values $(U, V)_r^- := (U_1, \ldots, U_{r-1}, V_1, \ldots, V_{r-1})$. This allows us to adaptively probe the boundary of G^*. A (randomized) active learning algorithm is defined by a sequence of conditional distributions $P_{U_r|(U,V)_r^-} := P_{U_r|(U_1,\ldots,U_{r-1},V_1,\ldots,V_{r-1})}$, which defines the joint distribution as follows:

$$P_{\mathbf{UV}} := \prod_{r=1}^{n} P_{U_r|(U,V)_r^-} P_{V_r|U_r} \tag{22}$$

where $P_{V_r|U_r} \equiv P_{V|U}$; in particular, conditioned on U_r, label V_r is independent of $(U_1, \ldots, U_{r-1})$. We assume that for each r, the conditional distribution $P_{U_r|(U,V)_r^-}$ has a density $p_{U_r|(U,V)_r^-}$ with respect to Lebesgue measure on $[0,1]^d$. Note that active learning algorithms correspond to channel coding with feedback, and to adaptive group testing algorithms [18]. Passive learning corresponds to channel coding without feedback, and to non-adaptive group testing algorithms.

We provide lower bounds on the excess risk of active learning algorithms that strengthen those in [10, Theorem 3], but our techniques can also be applied to [10, Theorem 4], which applies in the passive case. We use the packing set constructed in [10], which is defined via a hypercube class of joint distributions on (U, V). Fix an integer m (to be chosen later as a function of n). For each vector $\boldsymbol{\tau} \in \{0,1\}^{m^{d-1}}$, Castro and Nowak [10, Appendix C] construct a unique distribution of (U, V) whose feature conditional probability is denoted by $\eta_{\boldsymbol{\tau}}(u)$, and the corresponding Bayes classifier is denoted by $G^*_{\boldsymbol{\tau}}$. We denote this hypercube class of $2^{m^{d-1}}$ distributions by $\mathcal{F}_m$. Each distribution in $\mathcal{F}_m$ has the same U-marginal P_U. Thus the joint distribution is determined by the conditional distribution $P_{V|U}$. The conditional distributions in $\mathcal{F}_m$ (equivalently, the feature conditional probabilities $\eta_{\boldsymbol{\tau}}(u)$ for each $\boldsymbol{\tau} \in \{0,1\}^{m^{d-1}}$) are not explicitly defined here, but the definition can be found in the displayed equation at the foot of [10, p.2350]. The definition ensures that the hypercube class $\mathcal{F}_m \subseteq \mathsf{BF}(\alpha, \kappa, L, c)$.

The packing set defined in [10, Appendix C] is a subset of distributions in $\mathcal{F}_m$.

Remark 4.1 (Packing set construction)**.** *[10, Lemma 2] There exists a subset $\mathcal{A} \subseteq \{0,1\}^{m^{d-1}}$ of size $M+1$ with $M \geq 2^{m^{d-1}/8}$, whose elements have minimum pairwise Hamming distance at least $m^{d-1}/8$. It is then shown in [10] that this results in a packing set of functions $\mathcal{P}_{M+1,\beta_m/8} = \{\eta_{\boldsymbol{\tau}} : \boldsymbol{\tau} \in \mathcal{A}\}$, where $\beta_m = LHm^{-\alpha}$, and $\eta_{\boldsymbol{\tau}}(1|u) = P_{V|U}(u)$. (Hence $1 - \eta_{\boldsymbol{\tau}}(u) = P_{V|U}(0|u)$.) Here $\beta_m/8$ is a lower bound on the set distance between distinct elements of $\mathcal{P}_{M+1,\beta_m/8}$, defined as*

$$d_\Delta(G^*_{\boldsymbol{\tau}}, G^*_{\boldsymbol{\tau}'}) = \int \mathbb{I}\left(u \in (G^*_{\boldsymbol{\tau}} \Delta G^*_{\boldsymbol{\tau}'})\right) du, \tag{23}$$

and $H = \|h\|_1$ is the norm of a suitable smooth function h.

Furthermore, $\mathcal{P}_{M+1,\beta_m/8}$ contains the function $\eta_{\mathbf{0}}$, corresponding to the point $\boldsymbol{\tau} = (0, 0, \ldots, 0)$ in the hypercube. We use the other M functions in the packing set $\mathcal{P}_{M+1,\beta_m/8}$ to act as the M codewords $\{\theta_1, \ldots, \theta_M\}$ in Theorem 1. The Bayes classifiers corresponding to these codewords are denoted by $G^*_1, \ldots, G^*_M$.

As in Section 3, we prove an explicit bound on the bracketed term (8) in Theorem 1, with $(\mathbf{u}, \mathbf{v})$ corresponding to $\mathbf{y}$ in (8).

Lemma 4.1. *For an active learning algorithm described by $\prod_{r=1}^n P(U_r|(U,V)_r^-)$, we take $Q_{\mathbf{U},\boldsymbol{V}}(\boldsymbol{U}, \boldsymbol{V}) := \prod_{r=1}^n P(U_r|(U,V)_r^-) \prod_{r=1}^n P_{\mathbf{0}}(V_r|U_r)$, where $P_{\mathbf{0}}(V_r|U_r)$ is the conditional probability mass function determined by $\eta_{\mathbf{0}}$ which corresponds to the point $\boldsymbol{\tau} = (0, 0, \ldots, 0)$ in the hypercube.*

Further, for each $\boldsymbol{\tau} \in \mathcal{A}$ and $\boldsymbol{\tau} \neq \mathbf{0}$, we can take

$$P_{\boldsymbol{\tau}}(\mathbf{U}, \mathbf{V}) := \prod_{r=1}^n P(U_r|(U,V)_r^-) \prod_{r=1}^n P_{\boldsymbol{\tau}}(V_r|U_r),$$

where $P_{\boldsymbol{\tau}}(V_r|U_r)$ is the conditional probability mass function determined by $\eta_{\boldsymbol{\tau}}$.

Then, for any $\lambda > 0$, the bracketed term in (8) satisfies

$$\sum_{\boldsymbol{\tau}\in\mathcal{A},\boldsymbol{\tau}\neq\mathbf{0}} \frac{1}{M}\int_{\mathcal{Y}_n}\left(\frac{dP_{\boldsymbol{\tau}}}{dQ_{\mathbf{U},\mathbf{V}}}(\mathbf{u},\mathbf{v})\right)^{1+\lambda} dQ_{\mathbf{U},\mathbf{V}}(\mathbf{u},\mathbf{v}) \leq \exp\left(\frac{16c^2\beta_m^{2(\kappa-1)}\lambda n}{(1-2c\beta_m)}\right). \tag{24}$$

where for brevity we write an integral to represent integration and summation over the product space $\mathcal{Y}_n = [0,1]^{d\times n}\otimes\{0,1\}^n$, and $\beta_m = LHm^{-\alpha}$ as defined in Remark 4.1.

Proof. See Appendix D. □

Combining Lemma 4.1 with Theorem 1, we deduce the following lower bound.

Proposition 4.2. *Let $\rho = (d-1)/\alpha$. For any positive constant ν, the risk of a classifier $\widehat{G}_n$ learnt via any active learning algorithm satisfies*

$$\sup_{P_{UV}\in\mathsf{BF}(\alpha,\kappa,L,c)}\left\{\mathbb{E}[R(\widehat{G}_n)] - R(G^*)\right\} \geq \frac{4c\nu^{\kappa\alpha}}{\kappa}\left(\frac{LH}{32}\right)^{\kappa} n^{-\frac{\kappa}{2\kappa-2+\rho}}\,\epsilon_M, \tag{25}$$

where

$$\epsilon_M \geq 1 - \frac{1+\lambda}{\lambda^{\lambda/(1+\lambda)}}\exp\left(\frac{-\lambda n^{\frac{\rho}{2\kappa-2+\rho}}}{(1+\lambda)\nu^{d-1}}\left(\frac{\log 2}{8} - \frac{16c^2(LH)^{2\kappa-2}\nu^{d-1+2\alpha(\kappa-1)}}{1-2cLHn^{-1/(2\kappa-2+\rho)}/\nu}\right)\right). \tag{26}$$

Therefore, for large n we have

$$\begin{aligned}&\sup_{P_{UV}\in\mathsf{BF}(\alpha,\kappa,L,c)}\left\{\mathbb{E}[R(\widehat{G}_n)] - R(G^*)\right\}\\ &\geq \left[\frac{4c}{\kappa 32^{\kappa}}\left(\frac{\log 2}{128c^2}\right)^{\frac{\kappa}{2\kappa-2+\rho}}(LH)^{\frac{\kappa\rho}{2\kappa-2+\rho}}\right] n^{-\frac{\kappa}{2\kappa-2+\rho}}(1-o(1)).\end{aligned} \tag{27}$$

Proof. Consider the M codewords chosen from the packing set $\mathcal{P}_{M+1,\beta_m/8}$, as described in Remark 4.1, with corresponding Bayes classifiers $G_1^*,\ldots,G_M^*$. The minimum pairwise set distance between these Bayes classifiers is at least $\beta_m/8$. Equating the minimum distance of the packing set given by $2A\psi_n$ (in Theorem 1) to $\beta_m/8$, taking $A=1$ we obtain $\psi_n = \beta_m/16 = LHm^{-\alpha}/16$.

Using Lemma 4.1 in Theorem 1, for any $\lambda > 0$ the average error probability ϵ_M can be bounded from below as

$$\epsilon_M \geq 1 - \frac{1+\lambda}{\lambda^{\lambda/(1+\lambda)}}M^{-\lambda/(1+\lambda)}\exp\left(\frac{16c^2\beta_m^{2(\kappa-1)}\lambda n}{(1-2c\beta_m)(1+\lambda)}\right) \tag{28}$$

$$\overset{(a)}{\geq} 1 - \frac{1+\lambda}{\lambda^{\lambda/(1+\lambda)}}\exp\left(\frac{\lambda}{1+\lambda}\left(\frac{16c^2\beta_m^{2(\kappa-1)}n}{(1-2c\beta_m)} - \frac{m^{d-1}\log 2}{8}\right)\right). \tag{29}$$

where inequality (a) is obtained using the fact that packing set of distributions $\mathcal{P}_{M+1,\beta_m/8}$ has $M \geq 2^{m^{d-1}/8}$, as described in Remark 4.1 above.

Now, consider any distribution $P_{UV} \in \mathsf{BF}(\alpha, \kappa, L, c)$ with Bayes classifier G^*. It is shown in [10, p.2351] that the event

$$\left\{ d_\Delta(\widehat{G}_n, G^*) \geq \psi_n \right\} \Rightarrow \left\{ R(\widehat{G}_n) - R(G^*) \geq \min\left(\frac{4c}{\kappa 2^\kappa} \psi_n^\kappa, \psi_n \right) \right\}.$$

Defining $f(\psi_n) := \min\left(\frac{4c}{\kappa 2^\kappa}\psi_n^\kappa, \psi_n\right)$, we therefore obtain the following chain of inequalities:

$$\begin{aligned}
\epsilon_M &\overset{(b)}{\leq} \sup_{P_{UV} \in \mathsf{BF}(\alpha,\kappa,L,c)} \mathbb{P}\left(d_\Delta(\widehat{G}_n, G^*) \geq \psi_n \right) \\
&\leq \sup_{P_{UV} \in \mathsf{BF}(\alpha,\kappa,L,c)} \mathbb{P}\left(R(\widehat{G}_n) - R(G^*) \geq f(\psi_n) \right) \\
&\leq \sup_{P_{UV} \in \mathsf{BF}(\alpha,\kappa,L,c)} \frac{\mathbb{E}[R(\widehat{G}_n)] - R(G^*)}{f(\psi_n)},
\end{aligned}$$

where inequality (b) follows from (3). Hence, using $\psi_n = LHm^{-\alpha}/16$, we have[1]

$$\begin{aligned}
&\sup_{P_{UV} \in \mathsf{BF}(\alpha,\kappa,L,c)} \mathbb{E}[R(\widehat{G}_n)] - R(G^*) \\
&\geq f(\psi_n)\epsilon_M \\
&= \frac{4c}{\kappa}\left(\frac{LHm^{-\alpha}}{32}\right)^\kappa \epsilon_M \\
&\geq \frac{4c}{\kappa}\left(\frac{LHm^{-\alpha}}{32}\right)^\kappa \left[1 - \frac{1+\lambda}{\lambda^{\lambda/(1+\lambda)}} \exp\left(\frac{\lambda}{1+\lambda}\left(\frac{16c^2 n \beta_m^{2(\kappa-1)}}{(1-2c\beta_m)} - \frac{m^{d-1}\log 2}{8} \right)\right)\right],
\end{aligned}$$

where the last inequality is obtained using (28). The result follows by taking $m = n^{\frac{1}{\alpha(2\kappa-2)+d-1}}/\nu$.

To obtain (27), we choose the supremum of ν such that $\epsilon_M \to 1$ as $n \to \infty$ in order to obtain the largest possible prefactor in (4). □

Remark 4.2. *The paper [10] derives Fano-type bounds in this setting: in particular, taking $m = n^{\frac{1}{\alpha(2\kappa-2)+d-1}}/\nu$, the computation in p.2351 of [10] together with Theorem 6 of that paper gives the same bound as (25), but with a looser lower bound on ϵ_M given by*

$$\epsilon_M \geq \left(1 - 2\xi - \sqrt{\frac{32\xi\nu^{d-1}}{\log 2}} n^{-\frac{\rho}{4(\kappa-1)+2\rho}} \right), \tag{30}$$

where $\xi = \frac{256}{\log 2} c^2 (LH)^{2\kappa-2}\nu$. For the bound (30) to be meaningful, we need $\xi < \frac{1}{2}$, which implies $\nu < \frac{\log 2}{512 c^2 (LH)^{2\kappa-2}}$. Again, Proposition 4.2 provides a strong converse, while (30) provides a weak one (ϵ_M bounded away from zero).

[1] As $m \gg 1$ and $\kappa \geq 1$, we assume for brevity that $f(\psi_n) = \frac{4c}{\kappa 2^\kappa}\psi_n^\kappa$. This is always true for sufficiently large m when $\kappa > 1$. However, if $\kappa = 1$ and $c > \frac{1}{2}$, then $f(\psi_n) = \psi_n$; however, the definition of c in [10, Eq. (9)] implies that c can be restricted to $(0, \frac{1}{2}]$ without loss of generality.

5. Application to compressed sensing

We now describe how Theorem 1 can give improved risk lower bounds in compressed sensing. The goal in compressed sensing is to estimate a sparse vector $\mathbf{x} \in \mathbb{R}^n$ from a measurement $\mathbf{y} \in \mathbb{R}^m$ of the form

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}. \tag{31}$$

Here $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the (known) measurement matrix, and $\mathbf{w} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_m)$ is the noise vector. Throughout this section $\|\mathbf{x}\|$ denotes the L_2 Euclidean norm of a vector $\mathbf{x}$, and $\|\mathbf{A}\|_F$ denotes the Frobenius norm of a matrix $\mathbf{A}$, defined by $\|\mathbf{A}\|_F^2 = \mathsf{Tr}\left(\mathbf{A}^\mathbf{T}\mathbf{A}\right)$. We assume that the signal $\mathbf{x}$ is k-sparse, by considering $\mathbf{x} \in \Sigma_k$, where

$$\Sigma_k := \left\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_0 \leq k, \|\mathbf{x}\| = 1\right\}.$$

In the pioneering works [6, 9, 12] of Candès, Donoho, Romberg, and Tao, among others, it was shown that under suitable assumptions on $\mathbf{A}$ and the sparsity level k, the signal could be efficiently estimated to a high degree of accuracy, even when $m \ll n$. For example, when $\mathbf{A}$ satisfies the Restricted Isometry Property [7], reconstruction techniques based on minimizing the L_1 norm produce an estimate $\hat{\mathbf{x}}$ which satisfies

$$\frac{1}{n}\left\|\mathbf{x} - \hat{\mathbf{x}}\right\|^2 \leq C_0 \frac{k\sigma^2}{m} \log n$$

with high probability, provided that m is of order at least $k \log(n/k)$ [3]. (C_0 is a universal positive constant.)

To complement these achievability results, several authors, e.g., [1, 8, 25, 34] have derived lower bounds on the minimax risk under various assumptions on $\mathbf{A}$ and $\mathbf{x}$. The minimax risk is defined as

$$\mathsf{M}^*(\mathbf{A}) := \inf_{\hat{\mathbf{x}}} \sup_{\mathbf{x} \in \Sigma_k} \mathbb{E}\left[\frac{1}{n}\left\|\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}\right\|^2\right], \tag{32}$$

We show how Theorem 1 can be used to obtain a strong converse, improving by a constant factor the lower bound on $\mathsf{M}^*(\mathbf{A})$ obtained using Fano's inequality by Candès and Davenport in [8]. Using the probabilistic method, [8] shows the existence of a packing set of well-separated vectors in Σ_k. To be specific:

Remark 5.1 (Packing set construction)**.** *[8, Lemma 2] There exists a subset $\mathcal{X} \subseteq \Sigma_k$ of size $M := |\mathcal{X}| = (n/k)^{k/4}$ whose elements $\mathbf{u}_i$ satisfy*

1. $\|\mathbf{u}_i\|^2 = 1$.
2. $\|\mathbf{u}_i - \mathbf{u}_j\|^2 \geq \frac{1}{2}$ *for all* $1 \leq i, j \leq M$ *such that* $i \neq j$.
3. $\left\|\frac{1}{M}\sum_{i=1}^M \mathbf{u}_i\mathbf{u}_i^T - \frac{1}{n}\mathbf{I}\right\|_{\text{op}} \leq \beta/n$. *Here β is a constant that can be made arbitrarily small with growing n.*

The set $\mathcal{X}$ gives a packing set $\mathcal{P}_{M, C/\sqrt{2}} := \{\theta_1, \ldots, \theta_M\}$ of codewords with minimum distance $\|\theta_i - \theta_j\| \geq \frac{C}{\sqrt{2}}$, simply by taking $\theta_i = C\mathbf{u}_i$, where the value of C will be specified later.

In fact, we consider a subset of the packing set $\mathcal{P}_{M,C/\sqrt{2}}$, defined as follows:

Lemma 5.1. *Let $\delta_M \in [\frac{1}{M}, 1-\frac{1}{M}]$, where $M = (n/k)^{k/4}$. Then there exists a subset $\mathcal{P}_{M',C/\sqrt{2}} \subseteq \mathcal{P}_{M,C/\sqrt{2}}$ such that $M' := \lceil \delta_M M \rceil$ and*

$$\max_{\theta_i \in \mathcal{P}_{M',C/\sqrt{2}}} \|\mathbf{A}\theta_i\|^2 \leq \frac{\|\mathbf{A}\|_F^2 \, C^2(1+\beta)}{n(1-\delta_M)}.$$

Proof. We first bound the average over the packing set $\mathcal{P}_{M,C/\sqrt{2}}$, given by $\frac{1}{M}\sum_{i=1}^{M} \|\mathbf{A}\theta_i\|^2$. Using steps similar to those in [8, p.320], we have

$$\begin{aligned}
\frac{1}{M}\sum_{i=1}^{M} \|\mathbf{A}\theta_i\|^2 &= \frac{1}{M}\sum_{i=1}^{M} \mathsf{Tr}\left(\mathbf{A}\theta_i\theta_i^T\mathbf{A}^T\right) \\
&= \mathsf{Tr}\left((\mathbf{A^T A})\frac{1}{M}\sum_{i=1}^{M}\theta_i\theta_i^T\right) \\
&\overset{(a)}{\leq} \mathsf{Tr}\left(\mathbf{A^T A}\right)\left\|\frac{C^2}{M}\sum_{i=1}^{M}\mathbf{u}_i\mathbf{u}_i^T\right\| \\
&\overset{(b)}{\leq} \|\mathbf{A}\|_F^2 \, C^2 \frac{(1+\beta)}{n}. \qquad (33)
\end{aligned}$$

In the above chain, step (a) holds because both $(\mathbf{A^T A})$ and $\sum_{i=1}^{M}\mathbf{u}_i\mathbf{u}_i^T/M$ are positive semi-definite. Step (b) is obtained using Property 3 of the packing set as defined in Remark 5.1.

We use the fact that if the average of M non-negative numbers $c_1 \leq c_2 \ldots \leq c_M$ is c, then $c_j \leq \frac{c}{1-(j-1)/M}$, for $1 \leq j \leq M$ (because otherwise the sum of the $(M-j+1)$ largest numbers will exceed Mc). The result then follows by picking M' elements of $\mathcal{P}_{M,C/\sqrt{2}}$ in increasing order of $\|\mathbf{A}\theta\|^2$, and calling this set $\mathcal{P}_{M',C/\sqrt{2}}$. □

As we restrict attention to the subset $\mathcal{P}_{M',C/\sqrt{2}}$ in the rest of this section, with mild abuse of notation, let us denote its elements by $\{\theta_1, \ldots, \theta_{M'}\}$. Also, let $\phi(\mathbf{y}; \boldsymbol{m}, \boldsymbol{\Sigma})$ denote the normal density in $\mathbb{R}^m$ with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{\Sigma}$. Then, with μ denoting the Lebesgue measure on $\mathcal{Y}$, from the measurement model (31), for any θ_i we have

$$\frac{dP_{\mathbf{Y}|\theta_i}}{d\mu}(\mathbf{y}) = \phi(\mathbf{y}; \mathbf{A}\theta_i, \sigma^2\mathbf{I}_m) = \prod_{r=1}^{m}\phi(y_r; \mathbf{A}_r^T\theta_i, \sigma^2), \qquad (34)$$

where the rth row of $\mathbf{A}$ is denoted by $\mathbf{A}_r^T \in \mathbb{R}^n$. Further, we choose

$$\frac{dQ_{\mathbf{Y}}}{d\mu}(\mathbf{y}) = \phi(\mathbf{y}; \mathbf{0}, \sigma^2\mathbf{I}_m) = \prod_{r=1}^{m}\phi(y_r; 0, \sigma^2), \qquad (35)$$

and prove the following bound for the integral in (8).

Lemma 5.2. *Let $P_{\mathbf{Y}|\theta_i}$ and $Q_{\mathbf{Y}}$ given by (34) and (35), respectively. Then for any $\lambda > 0$,*

$$\int_{\mathcal{Y}} \left(\frac{dP_{\mathbf{Y}|\theta_i}}{dQ_{\mathbf{Y}}}(\mathbf{y}) \right)^{1+\lambda} dQ_{\mathbf{Y}}(\mathbf{y}) = \exp\left(\frac{\lambda(1+\lambda)}{2\sigma^2} \|\mathbf{A}\theta_i\|^2 \right). \tag{36}$$

Hence, the bracketed term in (8) can be bounded by

$$\sum_{i=1}^{M'} \frac{1}{M'} \int_{\mathcal{Y}} \left(\frac{dP_{\mathbf{Y}|\theta_i}}{dQ_{\mathbf{Y}}}(\mathbf{y}) \right)^{1+\lambda} dQ_{\mathbf{Y}}(\mathbf{y}) \le \exp\left(\frac{\lambda(1+\lambda)}{2\sigma^2} \frac{\|\mathbf{A}\|_F^2 C^2(1+\beta)}{n(1-\delta_M)} \right). \tag{37}$$

Proof. See Appendix E. □

Combining Lemma 5.2 with Theorem 1, we deduce the following lower bound.

Proposition 5.3. *For any $\lambda > 0$, $\Delta \in (0,1)$, and $M = (n/k)^{k/4}$, we have*

$$\begin{aligned} \mathsf{M}^*(\mathbf{A}) &= \inf_{\hat{\mathbf{x}}} \sup_{x \in \Sigma_k} \mathbb{E}\left[\frac{1}{n} \|\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}\|^2 \right] \\ &\ge \frac{\sigma^2}{4\|\mathbf{A}\|_F^2} \left(\frac{k}{4} \log \frac{n}{k} - 1 \right) \frac{(1-\Delta)}{(1+\lambda)(1+\beta)} \epsilon_M, \end{aligned} \tag{38}$$

where

$$\epsilon_M \ge 1 - (1+\lambda) \left(\frac{(\log M) M^{-\Delta}}{\lambda} \right)^{\lambda/(1+\lambda)}. \tag{39}$$

Therefore, for large n we have

$$\mathsf{M}^*(\mathbf{A}) \ge \frac{\sigma^2}{4\|\mathbf{A}\|_F^2} \left(\frac{k}{4} \log \frac{n}{k} \right) (1 - o(1)). \tag{40}$$

Proof. To apply Theorem 1, we equate the minimum distance $C/\sqrt{2}$ of the packing subset $\mathcal{P}'_{M',C/\sqrt{2}}$ to $2A\psi_n$. Taking $A = 1$ gives $\psi_n = \frac{C}{2\sqrt{2}}$. Then, taking $w(t) = t^2$, we deduce that

$$\inf_{\hat{\mathbf{x}}} \sup_{\mathbf{x} \in \Sigma_k} \mathbb{E}\left[\|\hat{\mathbf{x}}(\mathbf{y}) - \mathbf{x}\|^2 \right] \ge \left(\frac{C}{2\sqrt{2}} \right)^2 \epsilon_M.$$

We can bound ϵ_M by using (37) of Lemma 5.2 in Theorem 1:

$$\begin{aligned} \epsilon_M &\ge 1 - \frac{(1+\lambda)}{(\lambda M')^{\lambda/(1+\lambda)}} \exp\left(\frac{\lambda \|\mathbf{A}\|_F^2 C^2(1+\beta)}{2n\sigma^2(1-\delta_M)} \right) \\ &\ge 1 - \frac{(1+\lambda)}{(\lambda \delta_M)^{\lambda/(1+\lambda)}} \exp\left(\lambda \left(\frac{\|\mathbf{A}\|_F^2 C^2(1+\beta)}{2n\sigma^2(1-\delta_M)} - \frac{\log M}{1+\lambda} \right) \right) \end{aligned}$$

Hence for any fixed λ we obtain (38) and (39) by choosing

$$C^2 = \frac{2n\sigma^2(1-\delta_M)\log M}{\|\mathbf{A}\|_F^2\,(1+\beta)(1+\lambda)}(1-\Delta),$$

with $\delta_M = 1/\log M$ and $\Delta \in (0,1)$. To obtain (40), we recall from Remark 5.1 that β can be chosen arbitrarily small as $n \to \infty$. Furthermore, λ, Δ can also be arranged to go 0 (at suitably slow rates) as $n \to \infty$. □

Remark 5.2. *The paper [8] uses Fano's inequality to derive the following bound:*

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

Comparing with Proposition 5.3, we see that our result improves the bound by a factor close to 8 for large n.

Appendix A: Proof of Lemma 2.1

For $(X,Y) \in \mathcal{X}\times\mathcal{Y}$ consider hypotheses $H_0:\ (X,Y)\sim Q$ and $H_1:\ (X,Y)\sim P$, where we assume that $P \ll Q$ so that the Radon-Nikodym derivative $\frac{dP}{dQ}$ exists.

The following lemma can be found in [24, Lemma 12.2] (also [23, eq. (102)]).

Lemma A.1. *For any randomized test T to distinguish between the above hypotheses, and $\gamma > 0$, we have*

$$P[T=1] - \gamma Q[T=1] \leq P\left[\frac{dP}{dQ} > \gamma\right]. \tag{41}$$

We note that the maximum of the left hand side of (41) (over all all tests T) is the E_γ divergence [19, 27]. We use (41) to complete the proof of Lemma 2.1:

Proof of Lemma 2.1. As in [23, Theorem 26], let ϵ_M and ϵ'_M denote the average error probabilities over channels $P_{\mathbf{Y}|\theta}$ and $Q_{\mathbf{Y}|\theta} = Q_{\mathbf{Y}}$, respectively, for a channel code with M equiprobable codewords. Given $(\theta, \mathbf{Y})$, the result [23, Theorem 26] describes a (sub-optimal) hypothesis test based on the channel decoder to distinguish between $H_0 : (\theta,\mathbf{Y}) \sim Q_{\theta\mathbf{Y}} = \pi_\theta Q_{\mathbf{Y}}$ and $H_1 : (\theta,\mathbf{Y}) \sim P_{\theta\mathbf{Y}} = \pi_\theta P_{\mathbf{Y}|\theta}$. Let $T \in \{0,1\}$ denote the output of this test. It is shown in the proof of that theorem that the probability of Type I error, i.e., $Q[T=1]$ is $1-\epsilon'_M$, and the probability of type II error, i.e., $P[T=0] = \epsilon_M$. Applying Lemma A.1 to this hypothesis test yields that for any $\gamma > 0$,

$$1 - \epsilon'_M \geq \frac{1}{\gamma}\left(1 - \epsilon_M - P_{\theta\mathbf{Y}}\left[\frac{dP}{dQ} > \gamma\right]\right). \tag{42}$$

We observe that when $Q_{\mathbf{Y}|\theta} = Q_{\mathbf{Y}}$, any channel decoder has average error probability $\epsilon'_M = \frac{M-1}{M}$. The result in (10) follows by substituting this value for ϵ'_M in (42). □

Appendix B: Recovering Fano's Inequality from Theorem 1

Here we show how to obtain Fano's inequality from Theorem 1. We first establish a general converse result involving mutual information (equation (48)), and then obtain Fano's inequality from it.

From the variational representation in (12), for any $\lambda, \gamma > 0$ we have

$$\begin{aligned} \frac{1}{M} &\geq \frac{(1-\epsilon_M)}{\gamma} - \frac{1}{\gamma^{1+\lambda}} \sum_{i=1}^{M} \frac{1}{M} \int_{\mathcal{Y}} \left(\frac{dP_{\mathbf{Y}|\theta_i}}{dQ_{\mathbf{Y}}}(\mathbf{y}) \right)^{1+\lambda} dQ_{\mathbf{Y}}(\mathbf{y}) \\ &= \frac{(1-\epsilon_M)}{\gamma} - \frac{1}{\gamma^{1+\lambda}} \left(\lambda \mathcal{H}_{1+\lambda}(P_{\theta\mathbf{Y}} || Q_{\theta\mathbf{Y}}) + 1 \right), \end{aligned} \tag{43}$$

where $H_{1+\lambda}(P\|Q) := \frac{1}{\lambda} \int \left(\left(\frac{dP}{dQ} \right)^{1+\lambda} - 1 \right) dQ$ is the Hellinger divergence of order $(1+\lambda)$ from distribution P to distribution Q. We note from (9) that the Rényi and Hellinger divergences of order $(1+\lambda)$ are invertible functions of one another. We use the following bound [27, Theorem 8] for the Hellinger divergence:

$$\mathcal{H}_{1+\lambda}(P\|Q) \leq \kappa(\lambda, t) D(P\|Q), \tag{44}$$

where

$$t := \operatorname{ess\,sup} \frac{dP}{dQ}(x, y), \text{ for } (x, y) \sim Q \quad \text{and} \quad \kappa(\lambda, t) = \frac{\lambda + t^{1+\lambda} - (1+\lambda)t}{\lambda(t \log t + 1 - t)}. \tag{45}$$

We choose

$$Q_{\mathbf{Y}} = \overline{P}_{\mathbf{Y}} = \sum_{j=1}^{M} \frac{1}{M} P_{\mathbf{Y}|\theta_j}. \tag{46}$$

With this $Q_{\mathbf{Y}}$, we have that $t \leq M$ since $\overline{P}_{\mathbf{Y}}(\mathcal{A}) \geq M^{-1} P_{\mathbf{Y}|\theta_j}(\mathcal{A})$ for all measurable sets $\mathcal{A}$. We also have

$$D(P_{\theta\mathbf{Y}} || Q_{\theta\mathbf{Y}}) = I(\theta; \mathbf{Y}) = \sum_{i=1}^{M} \frac{1}{M} D(P_{\mathbf{Y}|\theta_i} || \overline{P}_{\mathbf{Y}}). \tag{47}$$

Substituting in (44) and then in (43), we obtain

$$\frac{1}{M} \geq \frac{(1-\epsilon_M)}{\gamma} - \frac{1}{\gamma^{1+\lambda}} \left(\lambda \kappa(\lambda, t) I(\theta; \mathbf{Y}) + 1 \right).$$

Maximizing over $\gamma > 0$ yields

$$M \leq \frac{(1+\lambda)^{1+1/\lambda}}{\lambda (1-\epsilon_M)^{1+1/\lambda}} \left[\lambda \kappa(\lambda, t) I(\theta; \mathbf{Y}) + 1 \right]^{1/\lambda}.$$

Taking logs, for any $\lambda > 0$ we have

$$\log M \leq \left(1 + \frac{1}{\lambda} \right) \log \frac{1+\lambda}{1-\epsilon_M} - \log \lambda + \frac{1}{\lambda} \log(1 + c(\lambda, t) I(\theta; \mathbf{Y})), \tag{48}$$

where

$$c(\lambda,t)=\lambda\kappa(\lambda,t)=\frac{\lambda+t^{1+\lambda}-(1+\lambda)t}{t\log t+1-t}\leq\frac{t^{\lambda}}{\log t-1}\quad\text{if } t\geq e, \tag{49}$$

where the final inequality follows by direct comparison.

Hence, for a fixed $\lambda>0$ and $M\geq 3$, using (49) and $t\leq M$ in (48) gives

$$0\leq(1+\lambda)\log\frac{1+\lambda}{1-\epsilon_M}-\lambda\log\lambda+\log\left(M^{-\lambda}+\frac{I(\theta;\mathbf{Y})}{\log M-1}\right).$$

Or,

$$\log M\leq 1+I(\theta;\mathbf{Y})\left(\frac{\lambda^{\lambda}(1-\epsilon_M)^{1+\lambda}}{(1+\lambda)^{1+\lambda}}-M^{-\lambda}\right)^{-1}. \tag{50}$$

Finally, noting that λ can be chosen arbitrarily small, choose $\lambda=1/(\log M)^{\alpha}$ for some $\alpha\in(0,1)$ in (50). We therefore have

$$\log M\leq 1+\frac{I(\theta;\mathbf{Y})}{1-\epsilon_M}(1+o(1)).$$

Using the expression for $I(\theta;\mathbf{Y})$ in (47) and rearranging, we get

$$\epsilon_M\geq 1-\frac{\frac{1}{M}\sum_{i=1}^{M}D(P_{\theta_i}||\overline{P})}{\log M-1}(1+o(1)),$$

where $o(1)$ denotes a term that goes to zero with growing M. We have thus recovered Fano's inequality in (5) to within $o(1)$ terms.

Appendix C: Proof of Lemma 3.1

Proof of Lemma 3.1. Since $Q_{\mathbf{Y}}$ is the uniform measure and each $\frac{dP_{\mathbf{Y}|\theta_i}}{dQ_{\mathbf{Y}}}$ corresponds to an $f_{\boldsymbol{\tau}}^n$, for each value of i, we can express the relevant integral as

$$\int_{\mathcal{Y}}\left(\frac{dP_{\mathbf{Y}|\theta_i}}{dQ_{\mathbf{Y}}}(\mathbf{y})\right)^2 dQ_{\mathbf{Y}}(\mathbf{y})=\int_{[0,1]^n}f_{\boldsymbol{\tau}}^n(\mathbf{y})^2d\mathbf{y}=\left(\int_0^1 f_{\boldsymbol{\tau}}(y)^2dy\right)^n. \tag{51}$$

For any $\boldsymbol{\tau}$ we can express the bracketed term on the RHS of (51) as

$$\begin{aligned}\int_0^1 f_{\boldsymbol{\tau}}(y)^2dy&=\int_0^1\left(1+\sum_{j=0}^{m-1}\tau_j g_j(y)\right)^2dy\\&=1+2\sum_{j=0}^{m-1}\tau_j\int_0^1 g_j(y)dy+\sum_{j=0}^{m-1}\sum_{k=0}^{m-1}\tau_j\tau_k\int_0^1 g_j(y)g_k(y)dy\\&\overset{(a)}{=}1+\frac{c^2a}{m^4}.\end{aligned} \tag{52}$$

Here equality (a) is obtained from (14) and (15) since $g_j(y)g_k(y) \equiv 0$ for $j \neq k$, and

$$\int_0^1 g_j(y)dy = \frac{c}{m^2}\int_{j/m}^{(j+1)/m} g(my-j)dy = \frac{c}{m^2}\int_0^1 g(u)\frac{du}{m} = 0,$$
$$\int_0^1 g_j(y)^2dy = \frac{c^2}{m^4}\int_{j/m}^{(j+1)/m} g(my-j)^2dy = \frac{c^2}{m^4}\int_0^1 g(u)^2\frac{du}{m} = \frac{c^2 a}{m^5}.$$

The result follows on substituting (52) into (51) and using $(1+x)^n \leq \exp(xn)$ for any $x \in \mathbb{R}$. □

Appendix D: Proof of Lemma 4.1

Proof of Lemma 4.1. We can express the key ratio on the LHS of (24) as

$$\frac{dP_{\boldsymbol{\tau}}}{dQ_{\mathbf{U},\mathbf{V}}}(\mathbf{U},\mathbf{V}) = \prod_{r=1}^n \frac{dP_{\boldsymbol{\tau}}}{dP_{\mathbf{0}}}(V_r|U_r) = \prod_{r=1}^n \frac{P_{\boldsymbol{\tau}}(V_r|U_r)}{P_{\mathbf{0}}(V_r|U_r)}. \tag{53}$$

We note that $dP_{\mathbf{0}}(V_r|U_r) = P_{\mathbf{0}}(V_r|U_r)d\nu(V_r)$, where ν represents the counting measure on $\{0,1\}$. Then, using (53) we write the integral in (24) as

$$\int_{\mathcal{Y}_n}\left(\prod_{r=1}^n \frac{dP_{\boldsymbol{\tau}}}{dP_{\mathbf{0}}}(v_r|u_r)\right)^{1+\lambda}\prod_{r=1}^n dP(u_r|(u,u)_r^-)\prod_{r=1}^n dP_{\mathbf{0}}(v_r|u_r)$$
$$= \int_{\mathcal{Y}_{n-1}}\prod_{r=1}^{n-1}\left(\frac{dP_{\boldsymbol{\tau}}}{dP_{\mathbf{0}}}(v_r|u_r)\right)^{1+\lambda}\Big[I_n\Big]\prod_{r=1}^{n-1} dP(u_r|(u,u)_r^-)\prod_{r=1}^{n-1} dP_{\mathbf{0}}(v_r|u_r)$$

where the inner integral I_n can be written as

$$I_n := \int_{[0,1]^d}\left(\frac{[\eta_{\boldsymbol{\tau}}(u_n)]^{1+\lambda}}{[\eta_{\mathbf{0}}(u_n)]^{\lambda}} + \frac{[1-\eta_{\boldsymbol{\tau}}(u_n)]^{1+\lambda}}{[1-\eta_{\mathbf{0}}(u_n)]^{\lambda}}\right)dP_{U_n|(U,V)_n^-}(u_n|(u,v)_n^-)$$
$$= \int_{[0,1]^d}\exp\left(\lambda D_{1+\lambda}\left(P_{\boldsymbol{\tau}}(\cdot|u_n)||P_0(\cdot|u_n)\right)\right)dP_{U_n|(U,V)_n^-}(u_n|(u,v)_n^-), \tag{54}$$

where we use the fact that the Rényi divergence of order $(1+\lambda)$ between two Bernoulli random variables with parameters $\eta_{\boldsymbol{\tau}}(u_n)$ and $\eta_{\mathbf{0}}(u_n)$, respectively, is

$$D_{1+\lambda}\left(P_{\boldsymbol{\tau}}(\cdot|u_n)||P_0(\cdot|u_n)\right) = \frac{1}{\lambda}\log\left(\frac{[\eta_{\boldsymbol{\tau}}(u_n)]^{1+\lambda}}{[\eta_{\mathbf{0}}(u_n)]^{\lambda}} + \frac{[1-\eta_{\boldsymbol{\tau}}(u_n)]^{1+\lambda}}{[1-\eta_{\mathbf{0}}(u_n)]^{\lambda}}\right).$$

Recalling that $\beta_m = LHm^{-\alpha}$ and $u_n \in [0,1]^d$, let us denote the dth coordinate of u_n by $u_{n,d}$. The construction of the hypercube class of functions $\mathcal{F}_m$ in [10, p.2350] ensures that for any $\boldsymbol{\tau}, \boldsymbol{\tau}' \in \{0,1\}^{m^{d-1}}$, the following properties hold.

$$\eta_{\boldsymbol{\tau}}(u_n) = \eta_{\boldsymbol{\tau}'}(u_n), \quad \beta_m \leq u_{n,d} \leq 1, \tag{55}$$
$$\frac{1}{2} - c\beta_m \leq \eta_{\boldsymbol{\tau}}(u_n) \leq \frac{1}{2} + c\beta_m, \quad 0 \leq u_{n,d} \leq \beta_m, \tag{56}$$
$$|\eta_{\boldsymbol{\tau}}(u_n) - \eta_{\boldsymbol{\tau}'}(u_n)| \leq 2c\beta_m^{\kappa-1}, \quad \forall\, u_n \in [0,1]^d. \tag{57}$$

We now use the following bound on the Rényi divergence due to Verdú and Sason [26, Theorem 3] for $u_{n,d} \leq \beta_m$:

$$D_{1+\lambda}\left(P_{\boldsymbol{\tau}}(\cdot|u_n)||P_0(\cdot|u_n)\right) \leq \log\left(1+\frac{2\delta^2}{\min_{v\in\{0,1\}} P_0(v|u_n)}\right), \tag{58}$$

where $\delta := |\eta_{\boldsymbol{\tau}}(u_n) - \eta_{\mathbf{0}}(u_n)|$ is the total variation distance between $P_{\boldsymbol{\tau}}(\cdot|u_n)$ and $P_0(\cdot|u_n)$. Using (57) for an upper bound on δ, and (56) for a lower bound on the minimum of $P_0(\cdot|u_n)$, we have from (58),

$$D_{1+\lambda}\left(P_{\boldsymbol{\tau}}(\cdot|u_n)||P_0(\cdot|u_n)\right) \leq \log\left(1+\frac{8c^2\beta_m^{2(\kappa-1)}}{\frac{1}{2}-c\beta_m}\right).$$

Substituting this bound in (54) to bound I_n, we obtain using $1+x \leq e^x$ that

$$I_n \leq \left(1+\frac{8c^2\beta_m^{2(\kappa-1)}}{\frac{1}{2}-c\beta_m}\right)^{\lambda} \leq \exp\left(\frac{16c^2\beta_m^{2(\kappa-1)}\lambda}{1-2c\beta_m}\right).$$

The result follows by induction on n. □

Appendix E: Proof of Lemma 5.2

Recall from (34) and (35) that we take $dP_{\theta_i}(\mathbf{y}|\theta_i) = \prod_{r=1}^m \phi(y_r; \mathbf{A}_r^T\theta_i, \sigma^2)$ and $dQ_{\mathbf{Y}}(\mathbf{y}) = \prod_{r=1}^m \phi(y_r; 0, \sigma^2)$.

Proof of Lemma 5.2. For any $\lambda > 0$, we have

$$\begin{aligned}\int_{\mathcal{Y}} \left(\frac{dP_{\mathbf{Y}|\theta_i}}{dQ_{\mathbf{Y}}}(\mathbf{y})\right)^{1+\lambda} dQ_{\mathbf{Y}}(\mathbf{y}) &= \prod_{r=1}^m \int_{\mathbb{R}} \frac{[\phi(y_r; \mathbf{A}_r^T\theta_i, \sigma^2)]^{1+\lambda}}{[\phi(y_r; 0, \sigma^2)]^{\lambda}} dy_r \\ &\overset{(a)}{=} \prod_{r=1}^m \exp\left(\frac{\lambda(1+\lambda)}{2\sigma^2}\left(\mathbf{A}_r^T\theta_i\right)^2\right) \\ &= \exp\left(\frac{\lambda(1+\lambda)}{2\sigma^2}\|\mathbf{A}\theta_i\|^2\right).\end{aligned}$$

The equality in step (a) is obtained by completing the square inside an exponential, and recognizing the remaining term as a multiple of a normal density.

To obtain (37), we use Lemma 5.1 to bound $\|\mathbf{A}\theta_i\|^2$ on the RHS of (36) for each $\theta_i \in \mathcal{P}_{M',C/\sqrt{2}}$. □

Acknowledgements

The authors thank B. Nakiboğlu for pointing out the optimal choice of $Q_{\mathbf{Y}}$ given in Remark 2.3. They also thank the Alan Turing Institute for funding to attend the scoping workshop 'Information-Theoretic Foundations and Algorithms for Large-Scale Data Inference'. RV would like to acknowledge support from a Marie Curie Career Integration Grant (Grant Number 631489).

References

[1] Aeron, S., Saligrama, V. and Zhao, M. (2010). Information theoretic bounds for compressed sensing. *IEEE Trans. Inform. Theory* **56** 5111–5130.

[2] Assouad, P. (1983). Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique* **296** 1021–1024.

[3] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics* 1705–1732.

[4] Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields* **71** 271–291.

[5] Birgé, L. (2005). A new lower bound for multiple hypothesis testing. *IEEE Trans. Inform. Theory* **51** 1611–1615.

[6] Candes, E. J. and Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory* **51** 4203 - 4215.

[7] Candes, E. J. (2008). The Restricted Isometry Property and its implications for compressed sensing. *Comptes Rendus Mathematique* **346** 589–592.

[8] Candes, E. J. and Davenport, M. A. (2013). How well can we estimate a sparse vector? *Applied and Computational Harmonic Analysis* **34** 317–323.

[9] Candes, E. J., Romberg, J. and Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory* **52** 489–509.

[10] Castro, R. M. and Nowak, R. D. (2008). Minimax bounds for active learning. *IEEE Trans. Inform. Theory* **54** 2339–2353.

[11] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. John Wiley, New York.

[12] Donoho, D. L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52** 1289–1306.

[13] Giné, E. and Nickl, R. (2015). *Mathematical foundations of infinite-dimensional statistical models* **40**. Cambridge University Press.

[14] Guntuboyina, A. (2011). Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Trans. Inform. Theory* **57** 2386–2399.

[15] Hayashi, M. (2007). Error exponent in asymmetric quantum hypothesis testing and its application to classical-quantum channel coding. *Physical Review A* **76** 062301.

[16] Hayashi, M. and Nagaoka, H. (2003). General formulas for capacity of classical-quantum channels. *IEEE Trans. Inform. Theory* **49** 1753–1768.

[17] Ibragimov, I. A. and Khasminskii, R. Z. (1977). Estimation of infinite-dimensional parameter in Gaussian white noise. *Doklady Akademii Nauk SSSR* **236** 1053–1055.

[18] Johnson, O. T. (2017). Strong converses for group testing in the finite blocklength regime. *IEEE. Trans. Inform. Theory* **63** 5923-5933.

[19] Liu, J., Cuff, P. and Verdú, S. (2017). E_γ-Resolvability. *IEEE Trans. Inform. Theory* **63** 2629–2658.

[20] Massart, P. (2007). Concentration Inequalities and Model Selection. In

Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003 (J. Picard, ed.) Springer.

[21] Nagaoka, H. (2005). Strong converse theorems in quantum information theory. In *Proceedings of ERATO Workshop on Quantum Information Science 2001, Univ. Tokyo, Tokyo, Japan, September 6–8, 2001, Asymptotic Theory in Quantum Statistical Inference* (M. Hayashi, ed.). World Scientific.

[22] Nakiboğlu, B. (2017). The Augustin center and the sphere packing bound for memoryless channels. In *IEEE Int. Symp. Information Theory* 1401–1405.

[23] Polyanskiy, Y., Poor, H. V. and Verdú, S. (2010). Channel coding rate in the finite blocklength regime. *IEEE Trans. Inform. Theory* **56** 2307–2359.

[24] Polyanskiy, Y. and Wu, Y. Lecture Notes on Information Theory. Online: http://people.lids.mit.edu/yp/homepage/data/itlectures_v4.pdf.

[25] Raskutti, G., Wainwright, M. J. and Yu, B. (2011). Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inform. Theory* **57** 6976–6994.

[26] Sason, I. and Verdú, S. (2015). Upper bounds on the relative entropy and Rényi divergence as a function of total variation distance for finite alphabets. In *Proc. IEEE Inf. Theory Workshop-Fall* 214–218.

[27] Sason, I. and Verdú, S. (2016). f-divergence Inequalities. *IEEE Trans. Inform. Theory* **62** 5973–6006.

[28] Sason, I. and Verdú, S. (2018). Arimoto–Rényi conditional entropy and Bayesian M-ary hypothesis testing. *IEEE Trans. Inform. Theory* **64** 4–25.

[29] Sibson, R. (1969). Information radius. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **14** 149–160.

[30] Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics* 135–166.

[31] Tsybakov, A. B. (2009). *Introduction to nonparametric estimation.* Springer Series in Statistics. Springer, New York.

[32] Vazquez-Vilar, G., Campo, A. T., Guillén i Fàbregas, A. and Martinez, A. (2016). Bayesian M-ary Hypothesis Testing: The Meta-Converse and Verdú-Han Bounds Are Tight. *IEEE Trans. Inform. Theory* **62** 2324–2333.

[33] Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Annals of Statistics* 1564–1599.

[34] Ye, F. and Zhang, C.-H. (2010). Rate Minimaxity of the Lasso and Dantzig Selector for the ℓ_q-Loss in ℓ_r-Balls. *Journal of Machine Learning Research* **11** 3519–3540.

[35] Yu, B. (1997). Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam* 423–435. Springer.