An Introductory Guide to Fano's Inequality with Applications in Statistical Estimation

Jonathan Scarlett[1] and Volkan Cevher[2]

[1] Department of Computer Science & Department of Mathematics, National University of Singapore, Singapore

[2] Laboratory for Information and Inference Systems (LIONS), EPFL, Switzerland

scarlett@comp.nus.edu.sg, volkan.cevher@epfl.ch

Abstract

Information theory plays an indispensable role in the development of algorithm-independent impossibility results, both for communication problems and for seemingly distinct areas such as statistics and machine learning. While numerous information-theoretic tools have been proposed for this purpose, the oldest one remains arguably the most versatile and widespread: Fano's inequality. In this chapter, we provide a survey of Fano's inequality and its variants in the context of statistical estimation, adopting a versatile framework that covers a wide range of specific problems. We present a variety of key tools and techniques used for establishing impossibility results via this approach, and provide representative examples covering group testing, graphical model selection, sparse linear regression, density estimation, and convex optimization.

Contents

1 Introduction

The tremendous progress in large-scale statistical inference and learning in recent years has been spurred by both practical and theoretical advances, with strong interactions between the two: Algorithms that come with *a priori* performance guarantees are clearly desirable, if not crucial, in practical applications, and practical issues are indispensable in guiding the theoretical studies.

Complementary to performance bounds for specific algorithms, a key role is also played by algorithm-independent impossibility results, stating conditions under which one cannot hope to achieve a certain goal. Such results provide definitive benchmarks for practical methods, serve as certificates for near-optimality, and help guide the practical developments towards directions where the greatest improvements are possible.

Since its introduction in 1948, the field of information theory has continually provided such benefits for the problems of storing and transmitting data, and has accordingly shaped the design of practical communication systems. In addition, recent years have seen mounting evidence that the tools and methodology of information theory reach far beyond communication problems, and can provide similar benefits *within the entire data processing pipeline*.

While many information-theoretic tools have been proposed for establishing impossibility results, the oldest one remains arguably the most versatile and widespread: Fano's inequality [1]. This fundamental inequality is not only ubiquitous in studies of communication, but has been applied extensively in statistical inference and learning problems; several examples are given in Table 1.

When applying Fano's inequality to such problems, one typically encounters a number of distinct challenges compared to those found in communication problems. The goal of this chapter is to introduce the reader to some of the key tools and techniques, explain their interactions and connections, and provide several representative examples.

1.1 Overview of Techniques

Throughout the chapter, we consider the following statistical estimation framework, which captures a broad range of problems including the majority of those listed in Table 1:

- There exists an unknown parameter θ, known to lie in some set Θ (e.g., a subset of $\mathbb{R}^p$), that we would like to estimate.
- In the simplest case, the estimation algorithm has access to a set of *samples* $\mathbf{Y} = (Y_1, \ldots, Y_n)$ drawn from some joint distribution $P_\theta^n(\mathbf{y})$ parametrized by θ. More generally, the samples may be drawn from some joint distribution $P_{\theta,\mathbf{X}}^n(\mathbf{y})$ parametrized by $(\theta, \mathbf{X})$, where $\mathbf{X} = (X_1, \ldots, X_n)$ are *inputs* that are either known in advance or selected by the algorithm itself.
- Given knowledge of $\mathbf{Y}$, as well as $\mathbf{X}$ if inputs are present, the algorithm forms an estimate $\hat{\theta}$ of θ, with the goal of the two being "close" in the sense that some *loss function* $\ell(\theta, \hat{\theta})$ is small. When referring to this step of the estimation algorithm, we will use the terms *algorithm* and *decoder* interchangeably.

We will initially use the following simple running example to exemplify some of the key concepts, and then turn to detailed applications in Sections 4 and 6.

Example 1. (1-sparse linear regression) A vector parameter $\theta \in \mathbb{R}^p$ is known to have at most one non-zero

Sparse and low rank problems		**Other estimation problems**	
Problem	References	Problem	References
Group testing	[2, 3]	Regression	[12, 13]
Compressive sensing	[4, 5]	Density estimation	[13, 14]
Sparse Fourier transform	[6, 7]	Kernel methods	[15, 16]
Principal component analysis	[8, 9]	Distributed estimation	[17, 18]
Matrix completion	[10, 11]	Local privacy	[19]
Sequential decision problems		**Other learning problems**	
Problem	References	Problem	References
Convex optimization	[20, 21]	Graph learning	[26, 27]
Active learning	[22]	Ranking	[28, 29]
Multi-armed bandits	[23]	Classification	[30, 31]
Bayesian optimization	[24]	Clustering	[32]
Communication complexity	[25]	Phylogeny	[33]

Table 1: Examples of applications for which impossibility results have been derived using Fano's inequality.

entry, and we are given n linear samples of the form $\mathbf{Y} = \mathbf{X}\theta + \mathbf{Z}$,[1] where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a known input matrix, and $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ is additive Gaussian noise. In other words, the i-th sample Y_i is a noisy sample of $\langle X_i, \theta \rangle$, where $X_i \in \mathbb{R}^p$ is the transpose of the i-th row of $\mathbf{X}$. The goal is to construct an estimate $\hat{\theta}$ such that the squared distance $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$ is small.

This example is an extreme case of *k-sparse linear regression*, in which θ has at most $k \ll p$ non-zero entries, i.e., at most k columns of $\mathbf{X}$ impact the output. The more general k-sparse recovery problem will be considered in Section 6.1.

We seek to establish algorithm-independent impossibility results, henceforth referred to as *converse bounds*, in the form of lower bounds on the *sample complexity*, i.e., the number of samples n required to achieve a certain average target loss. The following aspects of the problem significantly impact this goal, and their differences are highlighted throughout the chapter:

- Discrete vs. continuous: Depending on the application, the parameter set Θ may be discrete or continuous. For instance, in the 1-sparse linear regression example, one may consider the case that θ is known to lie in a finite set $\Theta \subseteq \mathbb{R}^p$, or one may consider the general estimation of a vector in the set

$$\Theta = \{\theta \in \mathbb{R}^p \,:\, \|\theta\|_0 \leq 1\}, \tag{1}$$

  where $\|\theta\|_0$ is the number of non-zeros in θ.

- Minimax vs. Bayesian: In the minimax setting, one seeks a decoder that attains a small loss for any given $\theta \in \Theta$, whereas in the Bayesian setting, one considers the average performance under some prior distribution on θ. Hence, these two variations respectively consider the worst-case and average-case performance with respect to θ. We focus primarily on the minimax setting throughout the chapter, and further discuss Bayesian settings in Section 7.2.

- Choice of target goal: Naturally, the target goal can considerably impact the fundamental performance limits of an estimation problem. For instance, in discrete settings, it is common to consider exact recovery, requiring that $\hat{\theta} = \theta$ (i.e., the 0-1 loss $\ell(\theta, \hat{\theta}) = \mathbb{1}\{\hat{\theta} \neq \theta\}$), but it is also of interest to understand to what extent approximate recovery criteria make the problem easier.

[1] Throughout the chapter, we interchange tuple-based notations such as $\mathbf{X} = (X_1, \ldots, X_n)$, $\mathbf{Y} = (Y_1, \ldots, Y_n)$ with vector/matrix notation such as $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{Y} \in \mathbb{R}^n$.

- Non-adaptive vs. adaptive sampling: In settings consisting of an input $\mathbf{X} = (X_1, \ldots, X_n)$ as introduced above, one often distinguishes between the *non-adaptive* setting, in which $\mathbf{X}$ is specified prior to observing any samples, and the *adaptive* setting, in which a given input X_i can be designed based on the past inputs $(X_1, \ldots, X_{i-1})$ and samples $(Y_1, \ldots, Y_{i-1})$. It is of significant interest to understand to what extent the additional freedom of adaptivity impacts the performance.

With these variations in mind, we proceed by outlining the main steps in obtaining converse bounds for statistical estimation via Fano's inequality.

1.1.1 Step 1: Reduction to Multiple Hypothesis Testing

The *multiple hypothesis testing* problem is defined as follows: An index $V \in \{1, \ldots, M\}$ is drawn from a prior distribution P_V, and a sequence of samples $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is drawn from a probability distribution $P_{\mathbf{Y}|V}$ parametrized by V. The M possible conditional distributions are known in advance, and the goal is to identify the index V with high probability given the samples.

In Figure 1, we provide a general illustration of how an estimation problem can be reduced to multiple hypothesis testing, possibly with the added twist of including inputs $\mathbf{X} = (X_1, \ldots, X_n)$. Supposing for the time being that we are in the minimax setting, the idea is to construct a hard subset of parameters $\{\theta_1, \ldots, \theta_M\}$ that are difficult to distinguish given the samples. We then lower bound the worst-case performance by the average over this hard subset. As a concrete example, a good choice for the 1-sparse linear regression problem is to set $M = 2p$ and consider the set of vectors of the form

$$\theta = (0, \ldots, 0, \pm\epsilon, 0, \ldots, 0), \tag{2}$$

where $\epsilon > 0$ is a constant. Hence, the non-zero entry of θ has a given magnitude, which can be selected to our liking for the purpose of proving a converse.

We envision an index $V \in \{1, \ldots, M\}$ being drawn uniformly at random and used to select the corresponding parameter θ_V, and the estimation algorithm being run to produce an estimate $\hat{\theta}$. If the parameters $\{\theta_1, \ldots, \theta_M\}$ are not too close and the algorithm successfully produces $\hat{\theta} \approx \theta_V$, then we should be able to infer the index V from $\hat{\theta}$. This entire process can be viewed as a problem of multiple hypothesis testing, where the v-th hypothesis is that the underlying parameter is θ_v ($v = 1, \ldots, M$). With this reduction, we can deduce that if the algorithm performs well then the hypothesis test is successful; the contrapositive statement is then that *if the hypothesis test cannot be successful, then the algorithm cannot perform well.*

In the 1-sparse linear regression example, we find from (2) that distinct $\theta_j, \theta_{j'}$ must satisfy $\|\theta_j - \theta_{j'}\|_2 \geq \sqrt{2} \cdot \epsilon$. As a result, we immediately obtain from the triangle inequality that the following holds:

$$\text{If} \quad \|\hat{\theta} - \theta_v\|_2 < \frac{\sqrt{2}}{2} \cdot \epsilon, \quad \text{then} \quad \operatorname*{arg\,min}_{v'=1,\ldots,M} \|\hat{\theta} - \theta_{v'}\| = v. \tag{3}$$

In other words, if the algorithm yields $\|\hat{\theta} - \theta_v\|_2^2 < \frac{\sqrt{2}}{2}\epsilon$, then V can be identified as the index corresponding to the closest vector to $\hat{\theta}$. Thus, sufficiently accurate estimation implies success in identifying V.

Discussion. Selecting the hard subset $\{\theta_1, \ldots, \theta_M\}$ of parameters is often considered somewhat of an art. While the proofs of existing converse bounds may seem easy in hindsight when the hard subset is known, coming up with a suitable choice for a new problem usually requires some creativity and/or exploration. Despite this, there exist general approaches that have proved to been effective in a wide range problems, which we exemplify in Sections 4 and 6.

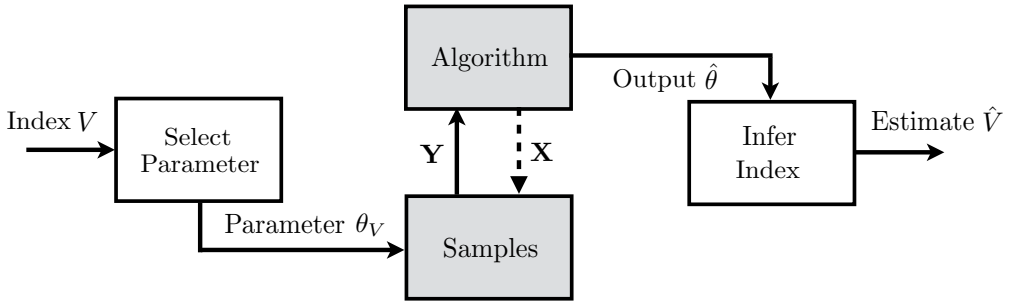

Figure 1: Reduction of minimax estimation to multiple hypothesis testing. The gray boxes are fixed as part of the problem statement, whereas the white boxes are constructed to our liking for the purpose of proving a converse bound. The dashed line marked with $\mathbf{X}$ is optional, depending on whether inputs are present.

In general, selecting the hard subset requires balancing conflicting goals: Increasing M so that the hypothesis test is more difficult, keeping the elements "close" so that they are difficult to distinguish, and keeping the elements "sufficiently distant" so that one can recover V from $\hat{\theta}$. Typically, one of the following three approaches is adopted: (i) explicitly construct a set whose elements are known or believed to be difficult to distinguish; (ii) prove the existence of such a set using probabilistic arguments; or (iii) consider packing as many elements as possible into the entire space. We will provide examples of all three kinds.

In the Bayesian setting, θ is already random, so we cannot use the above-mentioned method of lower bounding the worst-case performance by the average. Nevertheless, if Θ is discrete, we can still use the trivial reduction $V = \theta$ to form a multiple hypothesis testing problem with a possibly non-uniform prior. In the continuous Bayesian setting, one typically requires more advanced methods not covered in this chapter; we provide further discussion in Section 7.2.

1.1.2 Step 2: Application of Fano's Inequality

Once a multiple hypothesis test is set up, Fano's inequality provides a lower bound on its error probability in terms of the mutual information, which is one of the most fundamental information measures in information theory. The mutual information can often be explicitly characterized given the problem formulation, and a variety of useful properties are known for doing so, as outlined below.

We briefly state the standard form of Fano's inequality for the case that V is uniform on $\{1, \dots, M\}$:

$$\mathbb{P}[\hat{V} \neq V] \geq 1 - \frac{I(V;\hat{V}) + \log 2}{\log M}. \tag{4}$$

The intuition is as follows: The term $\log M$ represents the prior uncertainty (i.e., entropy) of V, and the mutual information $I(V;\hat{V})$ represents how much information $\hat{V}$ reveals about V. In order to have a small probability of error, we require that the information revealed is close to the prior uncertainty.

Beyond the standard form of Fano's inequality (4), it is useful to consider other variants, including approximate recovery and conditional versions. These are the topic of Section 2, and we discuss other alternatives in Section 7.2.

1.1.3 Step 3: Bounding the Mutual Information

In order to make lower bounds such as (4) explicit, we need to upper bound the mutual information therein. This often consists of tedious yet routine calculations, but there are cases where it is highly non-trivial. The

mutual information depends crucially on the choice of reduction in the first step.

The joint distribution of $(V, \hat{V})$ is decoder-dependent and usually very complicated, so to simplify matters, the typical first step is to apply an upper bound known as the data processing inequality. In the simplest case that there is no extra input to the sampling mechanism (i.e., $\mathbf{X}$ is absent in Figure 1), this inequality takes the form $I(V;\hat{V}) \leq I(V;\mathbf{Y})$ under the Markov chain $V \to \mathbf{Y} \to \hat{V}$. Thus, we are left to answer the question of how much information the samples reveal about the index V.

In Section 3, we introduce several useful tools for this purpose, including:

- Tensorization: If the samples $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are conditionally independent given V, we have $I(V;\mathbf{Y}) \leq \sum_{i=1}^{n} I(V;Y_i)$. Bounds of this type simplify the mutual information containing a set of observations to simpler terms containing only a single observation.

- KL divergence based bounds: Straightforward bounds on the mutual information reveal that if $\{P_{\theta_v}^n\}_{v=1,\ldots,M}$ are close in terms of KL divergence, then the mutual information is small. Results of this type are useful, as the relevant KL divergences can often be evaluated exactly or tightly bounded.

In addition to these, we introduce variations for cases that the input $\mathbf{X}$ is present in Figure 1, distinguishing between non-adaptive and adaptive sampling.

Toy example. To give a simple example of how this step is combined with the previous one, consider the case that we wish to identify one of M hypotheses, with the v-th hypothesis being that $\mathbf{Y} \sim P_v(\mathbf{y})$ for some distribution P_v on $\{0,1\}^n$. That is, the n observations $(Y_1, \ldots, Y_n)$ are binary-valued. Starting with the above-mentioned bound $I(V;\hat{V}) \leq I(V;\mathbf{Y})$, we simply write $I(V;\mathbf{Y}) \leq H(\mathbf{Y}) \leq n \log 2$, which follows since $\mathbf{Y}$ takes one of at most 2^n values. Substitution into (4) yields $P_{\mathrm{e}} \geq 1 - \frac{n+1}{\log_2 M}$, which means that achieving $P_{\mathrm{e}} \leq \delta$ requires $n \geq (1-\delta)\log_2 M - 1$. This formalizes the intuitive fact that reliably identifying one of $M \gg 1$ hypotheses requires roughly $\log_2 M$ binary observations.

2 Fano's Inequality and its Variants

In this section, we state various forms of Fano's inequality that will form the basis for the results in the remainder of the chapter.

2.1 Standard Version

We begin with the most simple and widely-used form of Fano's inequality. We use the generic notation V for the discrete random variable in a multiple hypothesis test, and we write its estimate as $\hat{V}$. In typical applications, one has a Markov chain relation such as $V \to \mathbf{Y} \to \hat{V}$, where $\mathbf{Y}$ is the collection of samples; we will exploit this fact in Section 3, but for now, one can think of $\hat{V}$ being randomly generated by any means given V.

The two fundamental quantities appearing in Fano's inequality are the conditional entropy $H(V|\hat{V})$, representing the uncertainty of V given its estimate, and the *error probability*:

$$P_{\mathrm{e}} = \mathbb{P}[\hat{V} \neq V]. \tag{5}$$

Since $H(V|\hat{V}) = H(V) - I(V;\hat{V})$, the conditional entropy is closely related to the mutual information, representing how much information $\hat{V}$ reveals about V.

Theorem 1. (Fano's inequality) *For any discrete random variables V and $\hat{V}$ on a common finite alphabet $\mathcal{V}$, we have*

$$H(V|\hat{V}) \leq H_2(P_\mathrm{e}) + P_\mathrm{e} \log\big(|\mathcal{V}| - 1\big), \tag{6}$$

where $H_2(\alpha) = \alpha \log \frac{1}{\alpha} + (1-\alpha) \log \frac{1}{1-\alpha}$ is the binary entropy function. In particular, if V is uniform on $\mathcal{V}$, we have

$$I(V;\hat{V}) \geq (1 - P_\mathrm{e}) \log |\mathcal{V}| - \log 2, \tag{7}$$

or equivalently,

$$P_\mathrm{e} \geq 1 - \frac{I(V;\hat{V}) + \log 2}{\log |\mathcal{V}|}. \tag{8}$$

Since the proof of Theorem 1 is widely accessible in standard references such as [34], we provide only an intuitive explanation of (6): To resolve the uncertainty in V given $\hat{V}$, we can first ask whether the two are equal, which bears uncertainty $H_2(P_\mathrm{e})$. In the case that they differ, which only occurs a fraction P_e of the time, the remaining uncertainty is at most $\log\big(|\mathcal{V}| - 1\big)$.

Remark 1. For uniform V, we obtain (7) by upper bounding $|\mathcal{V}| - 1 \leq |\mathcal{V}|$ and $H_2(P_\mathrm{e}) \leq \log 2$ in (6), and subtracting $H(V) = \log |\mathcal{V}|$ on both sides. While these additional bounds have a minimal impact for moderate to large values of $|\mathcal{V}|$, a notable case where one should use (6) is the binary setting, i.e., $|\mathcal{V}| = 2$. In this case, (7) is meaningless due to the right-hand side being negative, whereas (6) yields the following for uniform V:

$$I(V;\hat{V}) \geq \log 2 - H_2(P_\mathrm{e}). \tag{9}$$

It follows that the error probability is lower bounded as

$$P_\mathrm{e} \geq H_2^{-1}\big(\log 2 - I(V;\hat{V})\big), \tag{10}$$

where $H_2^{-1}(\cdot) \in \big[0, \frac{1}{2}\big]$ is the inverse of $H_2(\cdot) \in [0, \log 2]$ on the domain $\big[0, \frac{1}{2}\big]$.

2.2 Approximate Recovery

The notion of error probability considered in Theorem 1 is that of exact recovery, insisting that $\hat{V} = V$. More generally, one can consider notions of *approximate recovery*, where one only requires $\hat{V}$ to be "close" to V in some sense. This is useful for at least two reasons:

- Exact recovery is often a highly stringent criterion in discrete statistical estimation problems, and it is of considerable interest to understand to what extent moving to approximate recovery makes the problem easier;
- When we reduce continuous estimation problems to the discrete setting (*cf.*, Section 5), permitting approximate recovery will provide a useful additional degree of freedom.

We consider a general setup with a random variable V, an estimate $\hat{V}$, and an error probability of the form

$$P_\mathrm{e}(t) = \mathbb{P}\big[d(V,\hat{V}) > t\big] \tag{11}$$

for some real-valued function $d(v,\hat{v})$ and threshold $t \in \mathbb{R}$. In contrast to the exact recovery setting, there are interesting cases where V and $\hat{V}$ are defined on different alphabets, so we denote these by $\mathcal{V}$ and $\hat{\mathcal{V}}$, respectively.

One can interpret (11) as requiring $\hat{V}$ to be within a "distance" t of V. However, d need not be a true distance function, and need not even be symmetric nor take non-negative values. This definition of error probability in fact entails no loss of generality, since one can set $t = 0$ and $d(V, \hat{V}) = \mathbb{1}\{(V, \hat{V}) \in \mathcal{E}\}$ for an arbitrary set $\mathcal{E}$ containing the pairs that are considered errors.

In the following, we make use of the quantities

$$N_{\max}(t) = \max_{\hat{v} \in \hat{\mathcal{V}}} N_{\hat{v}}(t), \qquad N_{\min}(t) = \min_{\hat{v} \in \hat{\mathcal{V}}} N_{\hat{v}}(t), \tag{12}$$

where

$$N_{\hat{v}}(t) = \sum_{v \in \mathcal{V}} \mathbb{1}\{d(v, \hat{v}) \le t\} \tag{13}$$

counts the number of $v \in \mathcal{V}$ within a "distance" t of $\hat{v} \in \hat{\mathcal{V}}$.

Theorem 2. (Fano's inequality with approximate recovery) *For any random variables $V, \hat{V}$ on the finite alphabets $\mathcal{V}, \hat{\mathcal{V}}$, we have*

$$H(V|\hat{V}) \le H_2(P_{\mathrm{e}}(t)) + P_{\mathrm{e}}(t) \log \frac{|\mathcal{V}| - N_{\min}(t)}{N_{\max}(t)} + \log N_{\max}(t). \tag{14}$$

In particular, if V is uniform on $\mathcal{V}$, then

$$I(V; \hat{V}) \ge (1 - P_{\mathrm{e}}(t)) \log \frac{|\mathcal{V}|}{N_{\max}(t)} - \log 2, \tag{15}$$

or equivalently

$$P_{\mathrm{e}}(t) \ge 1 - \frac{I(V; \hat{V}) + \log 2}{\log \frac{|\mathcal{V}|}{N_{\max}(t)}}. \tag{16}$$

The proof is similar to that of Theorem 1, and can be found in [35].

By setting $d(v, \hat{v}) = \mathbb{1}\{v \ne \hat{v}\}$ and $t = 0$, we find that Theorem 2 recovers Theorem 1 as a special case. More generally, the bounds (15)–(16) resemble those for exact recovery in (7)–(8), but $\log |\mathcal{V}|$ is replaced by $\log \frac{|\mathcal{V}|}{N_{\max}(t)}$. When $\mathcal{V} = \hat{\mathcal{V}}$, one can intuitively think of the approximate recovery setting as dividing the space into regions of size $N_{\max}(t)$, and only requiring the correct region to be identified, thereby reducing the effective alphabet size to $\frac{|\mathcal{V}|}{N_{\max}(t)}$.

2.3 Conditional Version

When applying Fano's inequality, it is often useful to condition on certain random events and random variables. The following theorem states a general variant of Theorem 1 with such conditioning. Conditional forms for the case of approximate recovery (Theorem 2) follow in an identical manner.

Theorem 3. (Conditional Fano inequality) *For any discrete random variables V and $\hat{V}$ on a common alphabet $\mathcal{V}$, any discrete random variable A on an alphabet $\mathcal{A}$, and any subset $\mathcal{A}' \subseteq \mathcal{A}$, the error probability $P_{\mathrm{e}} = \mathbb{P}[\hat{V} \ne V]$ satisfies*

$$P_{\mathrm{e}} \ge \sum_{a \in \mathcal{A}'} \mathbb{P}[A = a] \frac{H(V|\hat{V}, A = a) - \log 2}{\log \big(|\mathcal{V}_a| - 1\big)}, \tag{17}$$

where $\mathcal{V}_a = \{v \in \mathcal{V} : \mathbb{P}[V = v \,|\, A = a] > 0\}$. For possibly continuous A, the same holds true with $\sum_{a \in \mathcal{A}'} \mathbb{P}[A = a](\cdots)$ replaced by $\mathbb{E}[\mathbb{1}\{A \in \mathcal{A}'\}(\cdots)]$.

Proof. We write $P_{\mathrm{e}} \geq \sum_{a \in \mathcal{A}'} \mathbb{P}[A = a]\mathbb{P}[\hat{V} \neq V \mid A = a]$, and lower bound the conditional error probability using Fano's inequality (*cf.*, Theorem 1) under the joint distribution of $(V, \hat{V})$ conditioned on $A = a$. □

Remark 2. Our main use of Theorem 3 will be to average over the input $\mathbf{X}$ (*cf.*, Figure 1) in the case that it is random and independent of V. In such cases, by setting $A = \mathbf{X}$ in (17) and letting $\mathcal{A}'$ contain all possible outcomes, we simply recover Theorem 1 with conditioning on $\mathbf{X}$ in the conditional entropy and mutual information terms. The approximate recovery version, Theorem 2, extends in the same way. In Section 4, we will discuss more advanced applications of Theorem 3, including (i) genie arguments, in which some information about V is revealed to the decoder, and (ii) typicality arguments, where we condition on V falling in some high-probability set.

3 Mutual Information Bounds

We saw in Section 2 that the mutual information $I(V; \hat{V})$ naturally arises from Fano's inequality when V is uniform. More generally, we have $H(V|\hat{V}) = H(V) - I(V; \hat{V})$, so we can characterize the conditional entropy by characterizing both the entropy and the mutual information. In this section, we provide some of the main useful tools for upper bounding the mutual information. For brevity, we omit the proofs of standard results commonly found in information theory textbooks, or simple variations thereof.

Throughout the section, the random variables V and $\hat{V}$ are assumed to be discrete, whereas the other random variables involved, including the inputs $\mathbf{X} = (X_1, \ldots, X_n)$ and samples $\mathbf{Y} = (Y_1, \ldots, Y_n)$, may be continuous. Hence, notation such as $P_Y(y)$ may represent either a probability mass function (PMF) or a probability density function (PDF).

3.1 Data Processing Inequality

Recall the random variables V, $\mathbf{X}$, $\mathbf{Y}$, and $\hat{V}$ in the multiple hypothesis testing reduction depicted in Figure 1. In nearly all cases, the first step in bounding a mutual information term such as $I(V; \hat{V})$ is to upper bound it in terms of the samples $\mathbf{Y}$, and possibly the inputs $\mathbf{X}$. By doing so, we remove the dependence on $\hat{V}$, and form a bound that is algorithm-independent.

The following lemma provides three variations along these lines. The three are all essentially equivalent, but are written separately since each will be more naturally suited to certain settings, as described below. Recall the terminology that $X \to Y \to Z$ *forms a Markov chain* if X and Z are conditionally independent given Y, or equivalently, Z depends on (X, Y) only through Y.

Lemma 1. (Data processing inequality)
(i) If $V \to \mathbf{Y} \to \hat{V}$ forms a Markov chain, then $I(V; \hat{V}) \leq I(V; \mathbf{Y})$.
(ii) If $V \to \mathbf{Y} \to \hat{V}$ forms a Markov chain conditioned on $\mathbf{X}$, then $I(V; \hat{V}|\mathbf{X}) \leq I(V; \mathbf{Y}|\mathbf{X})$.
(iii) If $V \to (\mathbf{X}, \mathbf{Y}) \to \hat{V}$ forms a Markov chain, then $I(V; \hat{V}) \leq I(V; \mathbf{X}, \mathbf{Y})$.

We will use the first part when $\mathbf{X}$ is absent or deterministic, the second part for random non-adaptive $\mathbf{X}$, and the third when the elements of $\mathbf{X}$ can be chosen adaptively based on the past samples (*cf.* Section 1.1).

3.2 Tensorization

One of the most useful properties of mutual information is *tensorization*: Under suitable conditional independence assumptions, mutual information terms containing length-n sequences (e.g., $\mathbf{Y} = (Y_1, \ldots, Y_n)$) can be

upper bounded by a sum of n mutual information terms, the i-th of which contains the corresponding entry of each associated vector (e.g., Y_i). Thus, we can reduce a complicated mutual information term containing sequences to a sum of simpler terms containing individual elements. The following lemma provides some of the most common scenarios in which such tensorization can be performed.

Lemma 2. (Tensorization of mutual information) *(i) If the entries of* $\mathbf{Y} = (Y_1, \ldots, Y_n)$ *are conditionally independent given* V*, then*

$$I(V; \mathbf{Y}) \leq \sum_{i=1}^{n} I(V; Y_i). \tag{18}$$

(ii) If the entries of $\mathbf{Y}$ *are conditionally independent given* $(V, \mathbf{X})$*, and* Y_i *depends on* $(V, \mathbf{X})$ *only through* (V, X_i)*, then*

$$I(V; \mathbf{Y}|\mathbf{X}) \leq \sum_{i=1}^{n} I(V; Y_i|X_i). \tag{19}$$

(iii) If, in addition to the assumptions in part (ii), Y_i *depends on* (V, X_i) *only through* $U_i = \psi_i(V, X_i)$ *for some deterministic function* ψ_i*, then*

$$I(V; \mathbf{Y}|\mathbf{X}) \leq \sum_{i=1}^{n} I(U_i; Y_i). \tag{20}$$

The proof is based on the sub-additivity of entropy, along with the conditional independence assumptions given. We will use the first part of the lemma when $\mathbf{X}$ is absent or deterministic, and the second and third parts for random non-adaptive $\mathbf{X}$. When $\mathbf{X}$ can be chosen adaptively based on the past samples (*cf.* Section 1.1), the following variant is used.

Lemma 3. (Tensorization of mutual information for adaptive settings) *(i) If* X_i *is a function of* (X_1^{i-1}, Y_1^{i-1})*, and* Y_i *is conditionally independent of* (X_1^{i-1}, Y_1^{i-1}) *given* (V, X_i)*, then*

$$I(V; \mathbf{X}, \mathbf{Y}) \leq \sum_{i=1}^{n} I(V; Y_i|X_i). \tag{21}$$

(ii) If, in addition to the assumptions in part (i), Y_i *depends on* (V, X_i) *only through* $U_i = \psi_i(V, X_i)$ *for some deterministic function* ψ_i*, then*

$$I(V; \mathbf{X}, \mathbf{Y}) \leq \sum_{i=1}^{n} I(U_i; Y_i). \tag{22}$$

The proof is based on the chain rule for mutual information, i.e., $I(V; \mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{n} I(X_i, Y_i; V \,|\, X_1^{i-1}, Y_1^{i-1})$, as well as suitable simplifications via the conditional independence assumptions.

Remark 3. The mutual information bounds in Lemma 3 are analogous to those used in the problem of communication with feedback [34, Sec. 7.12]. A key difference is that in the latter setting, the channel input X_i is a function of $(V, X_1^{i-1}, Y_1^{i-1})$, with V representing the message. In statistical estimation problems, the quantity V being estimated is typically unknown to the decision-maker, so the input X_i is only a function of (X_1^{i-1}, Y_1^{i-1})

Remark 4. Lemma 3 should be applied with care, since even if V is uniform on some set *a priori*, it may not be uniform conditioned on X_i. This is because in the adaptive setting, X_i depends on Y_1^{i-1}, which in turn depends on V.

3.3 KL Divergence Based Bounds

By definition, the mutual information is the KL divergence between the joint distribution and the product of marginals, $I(V;Y) = D(P_{VY} \| P_V \times P_Y)$, and can equivalently be viewed as a conditional divergence $I(V;Y) = D(P_{Y|V} \| P_Y | P_V)$. Viewing the mutual information in this way leads to a variety of useful bounds in terms of related KL divergence quantities, as the following lemma shows.

Lemma 4. (KL divergence based bounds) *Let P_V, P_Y, and $P_{Y|V}$ be the marginal distributions corresponding to a pair (V,Y), where V is discrete. For any auxiliary distribution Q_Y, we have*

$$I(V;Y) = \sum_v P_V(v) D\big(P_{Y|v}(\cdot \,|\, v) \,\big\|\, P_Y\big) \tag{23}$$

$$\leq \sum_v P_V(v) D\big(P_{Y|V}(\cdot \,|\, v) \,\big\|\, Q_Y\big) \tag{24}$$

$$\leq \max_v D\big(P_{Y|V}(\cdot \,|\, v) \,\big\|\, Q_Y\big), \tag{25}$$

and in addition,

$$I(V;Y) \leq \sum_{v,v'} P_V(v) P_V(v') D\big(P_{Y|V}(\cdot \,|\, v) \,\big\|\, P_Y(\cdot \,|\, v')\big) \tag{26}$$

$$\leq \max_{v,v'} D\big(P_{Y|V}(\cdot \,|\, v) \,\big\|\, P_{Y|V}(\cdot \,|\, v')\big). \tag{27}$$

Proof. We obtain (23) from the definition of mutual information, and (24) from the fact that $\mathbb{E}\big[\log \frac{P_{Y|V}(Y|V)}{P_Y(Y)}\big] = \mathbb{E}\big[\log \frac{P_{Y|V}(Y|V)}{Q_Y(Y)}\big] - \mathbb{E}\big[\log \frac{P_Y(Y)}{Q_Y(Y)}\big]$; the second term here is a KL divergence, and is therefore non-negative. We obtain (26) from (24) by noting that Q_Y can be chosen to be any of the $P_Y(\cdot \,|\, v')$, and the remaining inequalities (25) and (27) are trivial. □

The upper bounds in (24)–(27) are closely related, and often essentially equivalent in the sense that they lead to very similar converse bounds. In the authors' experience, it is usually slightly simpler to choose a suitable auxiliary distribution Q_Y and apply (25), rather than bounding the pairwise divergences as in (27). Examples will be given in Sections 4 and 6.

Remark 5. We have used the generic notation Y in Lemma 4, but in applications this may represent either the entire vector $\mathbf{Y}$, or a single one of its entries Y_i. Hence, the lemma may be used to bound $I(V;\mathbf{Y})$ directly, or one may first apply tensorization and then use the lemma to bound each $I(V;Y_i)$.

Remark 6. Lemma 4 can also be used to bound *conditional* mutual information terms such as $I(V;Y|X)$. Conditioned on any $X = x$, we can upper bound $I(V;Y|X=x)$ using Lemma 4, with an auxiliary distribution $Q_{Y|X=x}$ that may depend on x. For instance, doing this for (25) and then averaging over X, we obtain for any $Q_{Y|X}$ that

$$I(V;Y|X) \leq \max_v D\big(P_{Y|X,V}(\cdot \,|\, \cdot, v) \,\big\|\, Q_{Y|X} | P_X\big) \tag{28}$$

$$\leq \max_{x,v} D\big(P_{Y|X,V}(\cdot \,|\, x, v) \,\big\|\, Q_{Y|X}(\cdot \,|\, x)\big). \tag{29}$$

The bound (25) in Lemma 4 is useful when there exists a single auxiliary distribution Q_Y that is "close" to each $P_{Y|V}(\cdot|v)$ in KL divergence, i.e., $D\big(P_{Y|V}(\cdot \,|\, v) \,\big\|\, Q_Y\big)$ is small. It is natural to extend this idea by introducing multiple auxiliary distributions, and only requiring that any one of them is close to a given

$P_{Y|V}(\cdot|v)$. This can be viewed as "covering" the conditional distributions $\{P_{Y|V}(\cdot|v)\}_{v\in\mathcal{V}}$ with "KL divergence balls", and we will return to this viewpoint in Section 5.3.

Lemma 5. (Mutual information bound via covering) *Under the setup of Lemma 4, suppose there exist N distributions $Q_1(y), \dots, Q_N(y)$ such that for all v and some $\epsilon > 0$, it holds that*

$$\min_{j=1,\dots,N} D\big(P_{Y|V}(\cdot \,|\, v) \,\big\|\, Q_j\big) \le \epsilon. \tag{30}$$

Then we have

$$I(V;Y) \le \log N + \epsilon. \tag{31}$$

The proof is based on applying (24) with $Q_Y(y) = \frac{1}{N}\sum_{j=1}^N Q_j(y)$, and then lower bounding this summation over j by the value $j^*(v)$ achieving the minimum in (30). We observe that setting $N = 1$ in Lemma 5 simply yields (25).

3.4 Relations Between KL Divergence and Other Measures

As evidenced above, the KL divergence plays a crucial role in applications of Fano's inequality. In some cases, directly characterizing the KL divergence can still be difficult, and it is more convenient to bound it in terms of other divergences or distances. The following lemma gives a few simple examples of such relations; the reader is referred to [36] for a more thorough treatment.

Lemma 6. (Relations between divergence measures) *Fix two distributions P and Q, and consider the KL divergence $D(P\|Q) = \mathbb{E}_P\big[\log \frac{P(Y)}{Q(Y)}\big]$, total variation $d_{\mathrm{TV}}(P,Q) = \frac{1}{2}\mathbb{E}_Q\big[\big|\frac{P(Y)}{Q(Y)} - 1\big|\big]$, squared Hellinger distance $H^2(P,Q) = \mathbb{E}_Q\big[\big(\sqrt{\frac{P(Y)}{Q(Y)}} - 1\big)^2\big]$, and χ^2-divergence $\chi^2(P\|Q) = \mathbb{E}_Q\big[\big(\frac{P(Y)}{Q(Y)} - 1\big)^2\big]$. We have:*

- (KL vs. TV) *$D(P\|Q) \ge 2d_{\mathrm{TV}}(P,Q)^2$, whereas if P and Q are probability mass functions and each entry of Q is at least $\eta > 0$, then $D(P\|Q) \le \frac{2}{\eta} d_{\mathrm{TV}}(P,Q)^2$.*
- (Hellinger vs. TV) *$\frac{1}{2}H^2(P,Q) \le d_{\mathrm{TV}}(P,Q) \le H(P,Q)\sqrt{1 - \frac{H^2(P,Q)}{4}}$;*
- (KL vs. χ^2) *$D(P\|Q) \le \log(1 + \chi^2(P\|Q)) \le \chi^2(P\|Q)$.*

4 Applications – Discrete Settings

In this section, we provide two examples of statistical estimation problems in which the quantity being estimated is discrete: group testing and graphical model selection. Our goal is not to treat these problems comprehensively, but rather, to study particular instances that permit a simple analysis while still illustrating the key ideas and tools introduced in the previous sections. We consider the *high-dimensional* setting, in which the underlying number of parameters being estimated is much higher than the number of measurements. To simplify the final results, we will often write them using the asymptotic notation $o(1)$ for vanishing terms, but non-asymptotic variants are easily inferred from the proofs.

4.1 Group Testing

The group testing problem consists of determining a small subset of "defective" items within a larger set of items based on a number of pooled tests. A given test contains some subset of the items, and the binary

test outcome indicates, possibly in a noisy manner, whether or not *at least one* defective item was included in the test. This problem has a history in medical testing [37], and has regained significant attention following applications in communication protocols, pattern matching, database systems, and more.

In more detail, the setup is described as follows:

- In a population of p items, there are k unknown *defective items*. This defective set is denoted by $S \subseteq \{1, \dots, p\}$, and is assumed to be uniform on the set of $\binom{p}{k}$ subsets having cardinality k. Hence, in this example, we are in the Bayesian setting with a uniform prior. We focus on the sparse setting, in which $k \ll p$, i.e., defective items are rare.

- There are n tests specified by a *test matrix* $\mathbf{X} \in \{0,1\}^{n \times p}$: The (i,j)-th entry of $\mathbf{X}$, denoted by X_{ij}, indicates whether item j is included in test i. We initially consider the *non-adaptive* setting, where $\mathbf{X}$ is chosen in advance. We allow for this choice to be random; for instance, a common choice of random design is to let the entries of $\mathbf{X}$ be i.i.d. Bernoulli.

- To account for possible noise, we consider the following observation model:

$$Y_i = \left(\bigvee_{j \in S} X_{ij} \right) \oplus Z_i, \tag{32}$$

  where $Z_i \sim \text{Bernoulli}(\epsilon)$ for some $\epsilon \in \left[0, \frac{1}{2}\right)$, $\oplus$ denotes modulo-2 addition, and $\vee$ is the "OR" operation. In the channel coding terminology, this corresponds to passing the noiseless test outcome $\bigvee_{j \in S} X_{ij}$ through a binary symmetric channel. We assume that the noise variables Z_i are independent of each other and of $\mathbf{X}$, and we define the vector of test outcomes $\mathbf{Y} = (Y_1, \dots, Y_n)$.

- Given $\mathbf{X}$ and $\mathbf{Y}$, a decoder forms an estimate $\hat{S}$ of S. We initially consider the exact recovery criterion, in which the error probability is given by

$$P_{\text{e}} = \mathbb{P}[\hat{S} \neq S], \tag{33}$$

  where the probability with respect to S, $\mathbf{X}$, and $\mathbf{Y}$.

In the following subsections, we present several results and analysis techniques that are primarily drawn from [2, 3].

4.1.1 Exact Recovery with Non-Adaptive Testing

Under the exact recovery criterion (33), we have the following lower bound on the required number of tests. Recall that $H_2(\alpha) = \alpha \log \frac{1}{\alpha} + (1-\alpha) \log \frac{1}{1-\alpha}$ denotes the binary entropy function.

Theorem 4. (Group testing with exact recovery) *Under the preceding noisy group testing setup, in order to achieve $P_{\text{e}} \leq \delta$, it is necessary that*

$$n \geq \frac{k \log \frac{p}{k}}{\log 2 - H_2(\epsilon)} (1 - \delta - o(1)) \tag{34}$$

as $p \to \infty$, possibly with $k \to \infty$ simultaneously.

Proof. Since S is discrete-valued, we can use the trivial reduction to multiple hypothesis testing with $V = S$. Applying Fano's inequality (*cf.*, Theorem 1) with conditioning on $\mathbf{X}$ (*cf.*, Section 2.3), we obtain

$$I(S; \mathbf{Y}|\mathbf{X}) \geq (1-\delta) \log \binom{p}{k} - \log 2, \tag{35}$$

where we have also upper bounded $I(S; \hat{S}|\mathbf{X}) \leq I(S; \mathbf{Y}|\mathbf{X})$ using the data processing inequality (*cf.*, second part of Lemma 1), which in turn uses the fact that $S \to \mathbf{Y} \to \hat{S}$ conditioned on $\mathbf{X}$.

Let $U_i = \bigvee_{j \in S} X_{ij}$ denote the hypothetical noiseless outcome. Since the noise variables $\{Z_i\}_{i=1}^n$ are independent and Y_i depends on $(S, \mathbf{X})$ only through U_i (*cf.*, (32)), we can apply tensorization (*cf.*, third part of Lemma 2) to obtain

$$I(S; \mathbf{Y}|\mathbf{X}) \leq \sum_{i=1}^{n} I(U_i; Y_i) \tag{36}$$

$$\leq n\big(\log 2 - H_2(\epsilon)\big), \tag{37}$$

where (37) follows since Y_i is generated from U_i according to a binary symmetric channel, which has capacity $\log 2 - H_2(\epsilon)$. Substituting (37) and $\binom{p}{k} \geq \left(\frac{p}{k}\right)^k$ into (35) and rearranging, we obtain (34). □

Theorem 4 is known to be tight in terms of scaling laws whenever $\delta \in (0,1)$ is fixed and $k = o(p)$, and perhaps more interestingly, tight including constant factors as $\delta \to 0$ under the scaling $k = O(p^\theta)$ for sufficiently small $\theta > 0$. The matching achievability result in this regime can be proved using maximum-likelihood decoding [38]. However, achieving such a result using a computationally efficient decoder remains a challenging open problem.

4.1.2 Approximate Recovery with Non-Adaptive Testing

We now move to an approximate recovery criterion: The decoder outputs a list $\mathcal{L} \subseteq \{1, \dots, p\}$ of cardinality $L \geq k$, and we require that at least a fraction $(1-\alpha)k$ of the defective items appear in the list, for some $\alpha \in (0,1)$. It follows that the error probability can be written as

$$P_{\mathrm{e}}(t) = \mathbb{P}[d(S, \mathcal{L}) > t], \tag{38}$$

where $d(S, \mathcal{L}) = |S \backslash \mathcal{L}|$, and $t = \alpha k$. Notice that a higher value of L means more non-defective items may be included in the list, whereas a higher value of α means more defective items may be absent.

Theorem 5. (Group testing with approximate recovery) *Under the preceding noisy group testing setup with list size $L \geq k$, in order to achieve $P_{\mathrm{e}}(\alpha k) \leq \delta$ for some $\alpha \in (0,1)$ (not depending on p), it is necessary that*

$$n \geq \frac{(1-\alpha)k \log \frac{p}{L}}{\log 2 - H_2(\epsilon)}(1 - \delta - o(1)) \tag{39}$$

as $p \to \infty$, $k \to \infty$ and $L \to \infty$ simultaneously with $L = o(p)$.

Proof. We apply the approximate recovery version of Fano's inequality (*cf.*, Theorem 2) with $d(S, \mathcal{L}) = |S \backslash \mathcal{L}|$ and $t = \alpha k$ as above. For any $\mathcal{L}$ with cardinality L, the number of S with $d(S, \mathcal{L}) \leq \alpha k$ is given by $N_{\max}(t) = \sum_{j=0}^{\lfloor \alpha k \rfloor} \binom{p-L}{j}\binom{L}{k-j}$, which follows by counting the number of ways to place $k-j$ defective items in $\mathcal{L}$, and the remaining j defective items in the other $p-L$ entries. Hence, using Theorem 2 with conditioning

on $\mathbf{X}$ (*cf.*, Section 2.3), and applying the data processing inequality (*cf.*, second part of Lemma 1), we obtain

$$I(S;\mathbf{Y}|\mathbf{X}) \geq (1-\delta)\log\frac{\binom{p}{k}}{\sum_{j=0}^{\lfloor \alpha k \rfloor}\binom{p-L}{j}\binom{L}{k-j}} - \log 2. \tag{40}$$

By upper bounding the summation by $\lfloor \alpha k \rfloor + 1$ times the maximum value, and performing some asymptotic simplifications via the assumption $L = o(p)$, we can simplify the logarithm to $\big(k \log \frac{p}{L}\big)(1 + o(1))$ [39]. The theorem is then established by upper bounding the conditional mutual information using (37). □

Theorem 5 matches Theorem 4 up to the factor of $1-\alpha$ and the replacement of $\log \frac{p}{k}$ by $\log \frac{p}{L}$, suggesting that approximately recovery provides a minimal reduction in the number of tests even for moderate values of α and L. However, under approximate recovery, a near-matching achievability bound is known under the scaling $k = O(p^\theta)$ *for all* $\theta \in (0,1)$, rather than only sufficiently small θ [38].

4.1.3 Adaptive Testing

Next, we discuss the adaptive testing regime, in which a given input vector $X_i \in \{0,1\}^p$, corresponding to a single row of $\mathbf{X}$, is allowed to depend on the previous inputs and outcomes, i.e., $X_1^{i-1} = (X_1, \ldots, X_{i-1})$ and $Y_1^{i-1} = (Y_1, \ldots, Y_{i-1})$. In fact, it turns out that Theorems 4 and 5 still apply in this setting. Establishing this simply requires making the following modifications to the above analysis:

- Apply the data processing inequality in the form of the *third* part of Lemma 1, yielding (35) and (40) with $I(S;\mathbf{X},\mathbf{Y})$ in place of $I(X;\mathbf{Y}|\mathbf{X})$;
- Apply tensorization via Lemma 3 to deduce (36)–(37) with $I(S;\mathbf{X},\mathbf{Y})$ in place of $I(S;\mathbf{Y}|\mathbf{X})$.

In the regimes where Theorems 4 and/or 5 are known to have matching upper bounds with non-adaptive designs, we can clearly deduce that adaptivity provides no asymptotic gain. However, as with approximate recovery, adaptivity can significantly broaden the conditions under which matching achievability bounds are known, at least in the noiseless setting [40].

4.1.4 Discussion: General Noise Models

The preceding analysis can easily be extended to more general group testing models in which the observations $(Y_1, \ldots, Y_n)$ are conditionally independent given $\mathbf{X}$. A broad class of such models can be written in the form $(Y_i|N_i) \sim P_{Y|N}$, where $N_i = \sum_{j \in S} \mathbb{1}\{X_{ij} = 1\}$ denotes the number of defective items in the i-th test. In such cases, the preceding results hold true more generally when $\log 2 - H_2(\epsilon)$ is replaced by the capacity $\max_{P_N} I(N;Y)$ of the "channel" $P_{Y|N}$.

For certain models, we can obtain a better lower bound by applying a *genie argument*, along with the conditional form of Fano's inequality in Theorem 3. Fix $\ell \in \{1, \ldots, k\}$, and suppose that a uniformly random subset $S^{(1)} \subseteq S$ of cardinality $k - \ell$ is revealed to the decoder. This extra information can only make the group testing problem easier, so any converse bound for this modified setting remains valid for the original setting. Perhaps counter-intuitively, this idea can lead to a better final bound.

We only briefly outline the details of this more general analysis, and refer the interested reader to [3, 41]. Using Theorem 3 with $A = S^{(1)}$, and applying the data processing inequality and tensorization, one can obtain

$$P_{\mathrm{e}} \geq 1 - \frac{\sum_{i=1}^n I(N_i^{(0)}; Y_i | N_i^{(1)}) - \log 2}{\log\binom{p-k+\ell}{\ell}}, \tag{41}$$

where $N_i^{(1)} = \sum_{j \in S^{(1)}} \mathbb{1}\{X_{ij} = 1\}$, and $N_i^{(0)} = N_i - N_i^{(1)}$. The intuition is that we condition on $N_i^{(1)}$ since it is known via the genie, while the remaining information about Y_i is determined by $N_i^{(0)}$. Once (41) is established, it only remains to simplify the mutual information terms; see [3, 41] for further details.

4.2 Graphical Model Selection

Graphical models provide compact representations of the conditional independence relations between random variables, and frequently arise in areas such as image processing, statistical physics, computational biology, and natural language processing. The fundamental problem of *graphical model selection* consists of recovering the graph structure given a number of independent samples from the underlying distribution.

Graphical model selection has been studied under several different families of joint distributions, and also several different graph classes. We focus our attention on the commonly-used *Ising model* with binary observations, and on a simple graph class known as *forests*, defined to contain the graphs having no cycles.

Formally, the setup is described as follows:

- We are given n independent samples $Y_1, \ldots, Y_n$ from a p-dimensional joint distribution: $Y_i = (Y_{i1}, \ldots, Y_{ip})$ for $i = 1, \ldots, n$. This joint distribution is encoded by a graph $G = (V, E)$, where $V = \{1, \ldots, p\}$ is the *vertex set*, and $E \subseteq V \times V$ is the *edge set*. We use the terminology *vertex* and *node* interchangeably. We assume that there are no edges from a vertex to itself, and that the edges are *undirected*: $(i, j) \in E$ and $(j, i) \in E$ are equivalent, and only count as one edge.

- We focus on the *Ising model*, in which the observations are binary-valued, and the joint distribution of a given sample, say $Y_1 = (Y_{11}, \ldots, Y_{1p}) \in \{-1, 1\}^p$, is

$$P_G(y_1) = \frac{1}{Z} \exp\left(\lambda \sum_{(i,j) \in E} y_{1i} y_{1j}\right), \tag{42}$$

  where Z is a normalizing constant. Here $\lambda > 0$ is a parameter to the distribution dictating the edge strength; a higher value means it is more likely that $Y_{1i} = Y_{1j}$ for any given edge $(i, j) \in E$.

- We restrict the graph $G = (V, E)$ to be the set of all *forests*:

$$\mathcal{G}_{\text{forest}} = \big\{G : G \text{ has no cycles}\}, \tag{43}$$

  where a *cycle* is defined to be a path of distinct edges leading back to the start node, e.g., $(1, 4), (4, 2), (2, 1)$. A special case of a forest is a *tree*, which is an acyclic graph for which a path exists between any two nodes. One can view any forest as being a disjoint union of trees, each defined on some subset of V. See Figure 2 for an illustration.

- Let $\mathbf{Y} \in \{-1, 1\}^{n \times p}$ be the matrix whose i-th row contains the p entries of the i-th sample. Given $\mathbf{Y}$, a decoder forms an estimate $\hat{G}$ of G, or equivalently, an estimate $\hat{E}$ of E. We initially focus on the exact recovery criterion, in which the minimax error probability is given by

$$\mathcal{M}_n(\mathcal{G}_{\text{forest}}, \lambda) = \inf_{\hat{G}} \sup_{G \in \mathcal{G}_{\text{forest}}} \mathbb{P}_G[\hat{G} \neq G], \tag{44}$$

  where $\mathbb{P}_G$ denotes probability when the true graph is G, and the infimum is over all estimators.

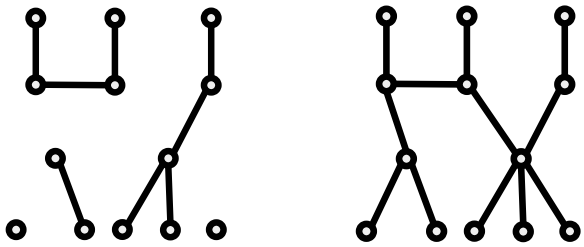

Figure 2: Two examples of graphs that are forests (i.e., acyclic graphs); the graph on the right is also a tree (i.e., a connected acyclic graph).

To our knowledge, Fano's inequality has not been applied previously in this exact setup; we do so using the general tools for Ising models given in [26, 27, 42, 43].

4.2.1 Exact Recovery

Under the exact recovery criterion, we have the following.

Theorem 6. (Exact recovery of forest graphical models) *Under the preceding Ising graphical model selection setup with a given edge parameter $\lambda > 0$, in order to achieve $\mathcal{M}_n(\mathcal{G}_{\text{forest}}, \lambda) \leq \delta$, it is necessary that*

$$n \geq \max\left\{\frac{\log p}{\log 2}, \frac{2\log p}{\lambda \tanh \lambda}\right\}(1 - \delta - o(1)) \tag{45}$$

as $p \to \infty$.

Proof. Recall from Section 1.1 that we can lower bound the worst-case error probability over $\mathcal{G}_{\text{forest}}$ by the average error probability over any subset of $\mathcal{G}_{\text{forest}}$. This gives us an important degree of freedom in the reduction to multiple hypothesis testing, and corresponds to selecting a hard subset $\theta_1, \ldots, \theta_M$ as described in Section 1.1.1. We refer to a given subset $\mathcal{G} \subseteq \mathcal{G}_{\text{forest}}$ as a *graph ensemble*, and provide two choices that lead to the two terms in (45).

For any choice of $\mathcal{G} \subseteq \mathcal{G}_{\text{forest}}$, Fano's inequality (Theorem 1) gives

$$n \geq \frac{(1-\delta)\log|\mathcal{G}| - \log 2}{I(G; Y_1)}, \tag{46}$$

for G uniform on $\mathcal{G}$, where we used $I(G; \hat{G}) \leq I(G; \mathbf{Y}) \leq nI(G; Y_1)$ by the data processing inequality and tensorization (*cf.* first parts of Lemmas 1 and 2).

Restricted ensemble 1: Let $\mathcal{G}_1$ be the set of all trees. It is well-known from graph theory that the number of trees on p nodes is $|\mathcal{G}_1| = p^{p-2}$ [44]. Moreover, since Y_1 is a length-p binary sequence, we have $I(G; Y_1) \leq H(Y_1) \leq p \log 2$. Hence, (46) yields $n \geq \frac{(1-\delta)(p-2)\log p - \log 2}{p \log 2}$, implying the first bound in (45).

Restricted ensemble 2: Let $\mathcal{G}_2$ be the set of graphs containing a single edge, so that $|\mathcal{G}_2| = \binom{p}{2}$. We will upper bound the mutual information using (25) in Lemma 4, choosing the auxiliary distribution Q_Y to be $P_{\overline{G}}$ with $\overline{G}$ being the empty graph. Thus, we need to bound $D(P_G \| P_{\overline{G}})$ for each $G \in \mathcal{G}_2$.

We first give an upper bound on $D(P_G \| P_{\overline{G}})$ for *any* two graphs $(G, \overline{G})$. We start with the trivial bound

$$D(P_G \| P_{\overline{G}}) \leq D(P_G \| P_{\overline{G}}) + D(P_G \| P_{\overline{G}}). \tag{47}$$

Recall the definition $D(P\|Q) = \mathbb{E}_P\big[\log \frac{P(Y)}{Q(Y)}\big]$, and consider the substitution of P_G and $P_{\overline{G}}$ according to

(42), with different normalizing constants Z_G and $Z_{\overline{G}}$. We see that when we sum the two terms in (47), the normalizing constants cancel, and we are left with

$$\begin{aligned} D(P_G \| P_{\overline{G}}) \leq & \sum_{(i,j) \in E \setminus \overline{E}} \lambda\big(\mathbb{E}_G[Y_{1i}Y_{1j}] - \mathbb{E}_{\overline{G}}[Y_{1i}Y_{1j}]\big) \\ & + \sum_{(i,j) \in \overline{E} \setminus E} \lambda\big(\mathbb{E}_{\overline{G}}[Y_{1i}Y_{1j}] - \mathbb{E}_G[Y_{1i}Y_{1j}]\big) \end{aligned} \tag{48}$$

for $G = (V, E)$ and $\overline{G} = (V, \overline{E})$.

In the case that G has a single edge (i.e., $G \in \mathcal{G}_2$) and $\overline{G}$ is the empty graph, we can easily compute $\mathbb{E}_{\overline{G}}[Y_{1i}Y_{1j}] = 0$, and (48) simplifies to

$$D(P_G \| P_{\overline{G}}) \leq \lambda \mathbb{E}_G[Y_{1i}Y_{1j}], \tag{49}$$

where (i, j) is the unique edge in G. Since Y_{1i} and Y_{1j} only take values in $\{-1, 1\}$, we have $\mathbb{E}_G[Y_{1i}Y_{1j}] = (+1)\mathbb{P}[Y_{1i} = Y_{1j}] + (-1)\mathbb{P}[Y_{1i} \neq Y_{1j}] = 2\mathbb{P}[Y_{1i} = Y_{1j}] - 1$, and letting E have a single edge in (42) yields $\mathbb{P}_G[(Y_{1i}, Y_{1j}) = (y_i, y_j)] = \frac{e^{\lambda y_i y_j}}{2e^{\lambda} + 2e^{-\lambda}}$, and hence $\mathbb{P}_G[Y_{1i} = Y_{1j}] = \frac{e^{\lambda}}{e^{\lambda} + e^{-\lambda}}$. Combining this with $\mathbb{E}_G[Y_{1i}Y_{1j}] = 2\mathbb{P}[Y_{1i} = Y_{1j}] - 1$ yields $\mathbb{E}_G[Y_{1i}Y_{1j}] = \frac{2e^{\lambda}}{e^{\lambda} + e^{-\lambda}} - 1 = \tanh \lambda$. Hence, using (49) along with (25) in Lemma 4, we obtain $I(G; Y_1) \leq \lambda \tanh \lambda$. Substitution into (46) (with $\log |\mathcal{G}| = (2 \log p)(1 + o(1))$) yields the second bound in (45). □

Theorem 6 is known to be tight up to constant factors whenever $\lambda = O(1)$ [44,45]: When λ is constant the lower bound becomes $n = \Omega(\log p)$, whereas for vanishing λ it simplifies to $n = \Omega\big(\frac{1}{\lambda^2} \log p\big)$.

4.2.2 Approximate Recovery

We consider the approximate recovery of $G = (V, E)$ with respect to the *edit distance* $d(G, \hat{G}) = |E \setminus \hat{E}| + |\hat{E} \setminus E|$, which is the number of edge additions and removals needed to transform G into $\hat{G}$ or vice versa. Since any forest can have at most $p - 1$ edges, it is natural to consider the case that an edit distance of up to αp is permitted, for some $\alpha > 0$. Hence, the minimax risk is given by

$$\mathcal{M}_n(\mathcal{G}_{\text{forest}}, \lambda, \alpha) = \inf_{\hat{G}} \sup_{G \in \mathcal{G}_{\text{forest}}} \mathbb{P}_G[d(G, \hat{G}) > \alpha p]. \tag{50}$$

In this setting, we have the following.

Theorem 7. (Approximate recovery of forest graphical models) *Under the preceding Ising graphical model selection setup with a given edge parameter $\lambda > 0$ and approximate recovery parameter $\alpha \in \big(0, \frac{1}{2}\big)$ (with the latter not depending on p), in order to achieve $\mathcal{M}_n(\mathcal{G}_{\text{forest}}, \lambda, \alpha) \leq \delta$, it is necessary that*

$$n \geq \max\left\{ \frac{(1 - \alpha) \log p}{\log 2}, \frac{2(1 - \alpha) \log p}{\lambda \tanh \lambda} \right\} (1 - \delta - o(1)) \tag{51}$$

as $p \to \infty$.

Proof. For any $\mathcal{G} \subseteq \mathcal{G}_{\text{forest}}$, Theorem 2 provides the following analog of (46):

$$n \geq \frac{(1 - \delta) \log \frac{|\mathcal{G}|}{N_{\max}(\alpha p)} - \log 2}{I(G; Y_1)} \tag{52}$$

for G uniform on $\mathcal{G}$, where $N_{\max}(t) = \max_{\hat{G}} \sum_{G \in \mathcal{G}} \mathbb{1}\{d(G, \hat{G}) \le t\}$ implicitly depends on $\mathcal{G}$. We again consider two restricted ensembles; the first is identical to the exact recovery setting, whereas the second is modified due to the fact that learning single-edge graphs with approximate recovery is trivial.

Restricted ensemble 1: Once again, let $\mathcal{G}_1$ be the set of all trees. We have already established $|\mathcal{G}_1| = (p-2)\log p$ and $I(G; Y_1) \le n \log 2$ for this ensemble, so it only remains to characterize $N_{\max}(\alpha p)$.

While the decoder may output a graph $\hat{G}$ not lying in $\mathcal{G}_1$, we can assume without loss of generality that $\hat{G}$ is always selected such that $d(\hat{G}, G^*) \le \alpha p$ for some $G^* \in \mathcal{G}_1$; otherwise, an error would be guaranteed. As a result, for any $\hat{G}$, and any $G \in \mathcal{G}_1$ such that $d(G, \hat{G}) \le \alpha p$, we have from the triangle inequality that $d(G, G^*) \le d(G, \hat{G}) + d(\hat{G}, G^*) \le 2\alpha p$, which implies that

$$N_{\max}(\alpha p) \le \sum_{G \in \mathcal{G}_1} \mathbb{1}\{d(G, G^*) \le 2\alpha p\}. \tag{53}$$

Now observe that since all graphs in $\mathcal{G}_1$ have exactly $p-1$ edges, transforming G to G^* requires removing j edges and adding j different edges, for some $j \le \alpha p$. Hence, we have

$$N_{\max}(\alpha p) \le \sum_{j=0}^{\lfloor \alpha p \rfloor} \binom{p-1}{j} \binom{\binom{p}{2} - p + 1}{j}. \tag{54}$$

By upper bounding the summation by $\lfloor \alpha p \rfloor + 1$ times the maximum, and performing some asymptotic simplifications, we can show that $\log N_{\max}(\alpha p) \le \big(\alpha p \log p\big)(1 + o(1))$. Substituting into (52) and recalling that $|\mathcal{G}_1| = (p-2)\log p$ and $I(G; Y_1) \le p \log 2$, we obtain the first bound in (51).

Restricted ensemble 2a: Let $\mathcal{G}_{2a}$ be the set of all graphs on p nodes containing exactly $\frac{p}{2}$ isolated edges; if p is an odd number, the same analysis applies with an arbitrary single node ignored. We proceed by characterizing $|\mathcal{G}_{2a}|$, $I(G; Y_1)$, and $N_{\max}(\alpha p)$. The number of graphs in the ensemble is $|\mathcal{G}_{2a}| = \binom{p}{2}\binom{p-2}{2} \cdots \binom{4}{2}\binom{2}{2} = \frac{p!}{2^{p/2}}$, and Stirling's approximation yields $\log |\mathcal{G}_{2a}| \ge \big(p \log p\big)(1 + o(1))$.

Since the KL divergence is additive for product distributions, and we established in the exact recovery case that the KL divergence between the distributions of a single-edge graph and an empty graph is at most $\lambda \tanh \lambda$, we deduce that $D(P_G \| P_{\overline{G}}) \le \frac{p}{2} \lambda \tanh \lambda$ for any $G \in \mathcal{G}_{2a}$, where $\overline{G}$ is the empty graph. We therefore obtain from Lemma 4 that $I(G; Y_1) \le \frac{p}{2} \lambda \tanh \lambda$.

A similar argument to that of Ensemble 1 yields $N_{\max}(\alpha p) \le \sum_{j=0}^{\lfloor \alpha p \rfloor} \binom{\frac{p}{2}}{j} \binom{\binom{p}{2} - \frac{p}{2}}{j}$, in analogy with (54). This again simplifies to $N_{\max}(\alpha p) \le \big(\alpha p \log p\big)(1 + o(1))$, and having established $\log |\mathcal{G}_{2a}| \ge \big(p \log p\big)(1 + o(1))$ and $I(G; Y_1) \le \frac{p}{2} \lambda \tanh \lambda$, substitution into (52) yields the second bound in (51). □

The bound in Theorem 7 matches that of Theorem 6 up to a multiplicative factor of $1 - \alpha$, thus suggesting that approximate recovery does not significantly help in reducing the required number of samples, at least in the minimax sense, for the Ising model and forest graph class.

4.2.3 Adaptive Sampling

We now return to the exact recovery setting, and consider a modification in which we have an added degree of freedom in the form of *adaptive sampling*:

- The algorithm proceeds in rounds; in round i, the algorithm queries a subset of the p nodes indexed by $X_i \in \{0, 1\}^p$, and the corresponding sample Y_i is generated as follows:

– The joint distribution of the entries of Y_i, corresponding to the entries where X_i is one, coincides with the corresponding marginal distribution of P_G, with independence between rounds;

– The values of the entries of Y_i, corresponding to the entries where X_i is zero, are given by $*$, a symbol indicating that the node was not observed.

We allow X_i to be selected based on the past queries and samples, namely, $X_1^{i-1} = (X_1, \ldots, X_{i-1})$ and $Y_1^{i-1} = (Y_1, \ldots, Y_{i-1})$.

- Let $n(X_i)$ denote the number of ones in X_i, i.e., the number of nodes observed in round i. While we allow the total number of rounds to vary, we restrict the algorithm to output an estimate $\hat{G}$ after observing at most n_{node} nodes. This quantity is related to n in the non-adaptive setting according to $n_{\text{node}} = np$, since in the non-adaptive setting we always observe all p nodes.

- The minimax risk is given by

$$\mathcal{M}_{n_{\text{node}}}(\mathcal{G}_{\text{forest}}, \lambda) = \inf_{\hat{G}} \sup_{G \in \mathcal{G}_{\text{forest}}} \mathbb{P}_G[\hat{G} \neq G], \tag{55}$$

where the infimum is over all adaptive algorithms that observe at most n_{node} nodes in total.

Theorem 8. (Adaptive sampling for forest graphical models) *Under the preceding Ising graphical model selection problem with adaptive sampling and a given parameter $\lambda > 0$, in order to achieve $\mathcal{M}_{n_{\text{node}}}(\mathcal{G}_{\text{forest}}, \lambda) \leq \delta$, it is necessary that*

$$n_{\text{node}} \geq \max\left\{\frac{p \log p}{\log 2}, \frac{2p \log p}{\lambda \tanh \lambda}\right\}(1 - \delta - o(1)) \tag{56}$$

as $p \to \infty$.

Proof. We prove the result using Ensemble 1 and Ensemble 2a above. We let N denote the number of rounds; while this quantity is allowed to vary, we can assume without loss of generality that $N = n_{\text{node}}$ by adding or removing rounds where no nodes are queried. For any subset $\mathcal{G} \subseteq \mathcal{G}_{\text{forest}}$, applying Fano's inequality (*cf.*, Theorem 1) and tensorization (*cf.*, first part of Theorem 3) yields

$$\sum_{i=1}^{N} I(G; Y_i | X_i) \geq (1 - \delta) \log |\mathcal{G}| - \log 2, \tag{57}$$

where G is uniform on $\mathcal{G}$.

Restricted ensemble 1: We again let $\mathcal{G}_1$ be the set of all trees, for which we know that $|\mathcal{G}| = p^{p-2}$. Since the $n(X_i)$ entries of Y_i differing from $*$ are binary, and those equaling $*$ are deterministic given X_i, we have $I(G; Y_i | X_i = x_i) \leq n(x_i) \log 2$. Averaging over X_i and summing over i yields $\sum_{i=1}^{N} I(G; Y_i | X_i) \leq \sum_{i=1}^{N} \mathbb{E}[n(X_i)] \log 2 \leq n_{\text{node}} \log 2$, and substitution into (57) yields the first bound in (56).

Restricted ensemble 2a: We again use the above-defined ensemble $\mathcal{G}_{2a}$ of graphs with $\frac{p}{2}$ isolated edges, for which we know that $|\mathcal{G}_{2a}| \geq \big(p \log p\big)(1 + o(1))$. In this case, when we observe $n(X_i)$ nodes, the subgraph corresponding to these observed nodes has at most $\frac{n(X_i)}{2}$ edges, all of which are isolated. Hence, using Lemma 4, the above-established fact that the KL divergence from a single-edge graph to the empty graph is at most $\lambda \tanh \lambda$, and the additivity of KL divergence for product distributions, we deduce that $I(G; Y_i | X_i = x_i) \leq \frac{n(x_i)}{2} \lambda \tanh \lambda$. Averaging over X_i and summing over i yields $\sum_{i=1}^{N} I(G; Y_i | X_i) \leq \frac{1}{2} n_{\text{node}} \lambda \tanh \lambda$, and substitution into (57) yields the second bound in (56). □

The threshold in Theorem 8 matches that of Theorem 6, and in fact, a similar analysis under approximate recovery also recovers the threshold in Theorem 7. This suggests that adaptivity is of limited help in the minimax sense for the Ising model and forest graph class. There are, however, other instances of graphical model selection where adaptivity provably helps [43, 46].

4.2.4 Discussion: Other Graph Classes

Degree and edge constraints: While the class $\mathcal{G}_{\text{forest}}$ is a relatively easy class to handle, similar techniques have also been used for more difficult classes, notably including those that place restrictions on the maximal degree d and/or the number of edges k. Ensembles 2 and 2a above can again be used, and the resulting bounds are tight in certain scaling regimes where $\lambda \to 0$, but loose in other regimes due to their lack of dependence on k and d. To obtain bounds with such a dependence, alternative ensembles have been proposed consisting of sub-graphs with highly correlated nodes [26, 27, 42].

For instance, suppose that a group of $d+1$ nodes has all possible edges connected except one. Unless d or the edge strength λ are small, the high connectivity makes the nodes very highly correlated, and the sub-graph is difficult to distinguish from a fully-connected sub-graph. This is in contrast with Ensembles 2 and 2a above, whose graphs are difficult to distinguish from the empty graph.

Bayesian setting: Beyond minimax estimation, it is also of interest to understand the fundamental limits of random graphs. A particularly prominent example is the Erdös-Rényi random graph, in which each edge is independently included with some probability $q \in (0,1)$. This is a case where the conditional form of Fano's inequality has proved useful; specifically, one can apply Theorem 3 with $A = G$, and $\mathcal{A}$ equal to the following *typical set* of graphs:

$$\mathcal{T} = \left\{ G : (1-\epsilon) q \binom{p}{2} \le |E| \le (1+\epsilon) q \binom{p}{2} \right\}, \tag{58}$$

where $\epsilon > 0$ is a constant. Standard properties of typical sets [34] yield that $\mathbb{P}[G_{\text{ER}} \in \mathcal{T}] \to 1$, $|\mathcal{T}| = e^{\left(H_2(q)\binom{p}{2}\right)(1+O(\epsilon))}$, and $H(V|V \in \mathcal{T}) = \left(H_2(q)\binom{p}{2}\right)(1+O(\epsilon))$ whenever $q\binom{p}{2} \to \infty$, and once these facts are established, Theorem 3 yields the following following necessary condition for $P_{\text{e}} \le \delta$:

$$n \ge \frac{p H_2(q)}{2 \log 2} (1 - \delta - o(1)). \tag{59}$$

For instance, in the case that $q = O\left(\frac{1}{p}\right)$ (i.e., there are $O(p)$ edges on average), we have $H_2(q) = \Theta\left(\frac{\log p}{p}\right)$, and we find that $n = \Omega(\log p)$ samples are necessary. This scaling is tight when λ is constant [45], whereas improved bounds for other scalings can be found in [27].

5 From Discrete to Continuous

Thus far, we have focused on using Fano's inequality to provide converse bounds for the estimation of discrete quantities. In many, if not most, statistical applications, one is instead interested in estimating continuous quantities; examples include linear regression, covariance estimation, density estimation, and so on. It turns out that the discrete form of Fano's inequality is still broadly applicable in such settings. The idea, as outlined in Section 1, is to choose a finite subset that still captures the inherent difficulty in the problem. In this section, we present several tools used for this purpose.

5.1 Minimax Estimation Setup

Recall the setup described in Section 1.1: A parameter θ is known to lie in some subset Θ of a continuous domain (e.g., $\mathbb{R}^p$), the samples $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are drawn from a joint distribution $P_\theta^n(\mathbf{y})$, an estimate $\hat{\theta}$ is formed, and the loss incurred is $\ell(\theta, \hat{\theta})$. For clarity of exposition, we focus primarily on the case that there is no input, i.e., $\mathbf{X}$ in Figure 1 is absent or deterministic. However, the main results (*cf.*, Theorems 9 and 10 below) extend to settings with inputs as described in Section 1.1; the mutual information $I(V; \mathbf{Y})$ is replaced by $I(V; \mathbf{Y}|\mathbf{X})$ in the non-adaptive setting, or $I(V; \mathbf{X}, \mathbf{Y})$ in the adaptive setting.

In continuous settings, the reduction to multiple hypothesis testing (*cf.*, Figure 1) requires that the loss function is sufficiently well-behaved. We focus on a widely-considered class of functions that can be written as

$$\ell(\theta, \hat{\theta}) = \Phi\big(\rho(\theta, \hat{\theta})\big), \tag{60}$$

where $\rho(\theta, \theta')$ is a metric, and $\Phi(\cdot)$ is an increasing function from $\mathbb{R}_+$ to $\mathbb{R}_+$. For instance, the squared-ℓ_2 loss $\ell(\theta, \theta') = \|\theta - \theta'\|_2^2$ clearly takes this form.

We focus on the minimax setting, defining the *minimax risk* as follows:

$$\mathcal{M}_n(\Theta, \ell) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta\big[\ell(\theta, \hat{\theta})\big], \tag{61}$$

where the infimum is over all estimators $\hat{\theta} = \hat{\theta}(\mathbf{Y})$, and $\mathbb{E}_\theta$ denotes expectation when the underlying parameter is θ. We subsequently define $\mathbb{P}_\theta$ analogously.

5.2 Reduction to the Discrete Case

We present two related approaches to reducing the continuous estimation problem to a discrete one. The first, based on the standard form of Fano's inequality in Theorem 1, was discovered much earlier [12], and accordingly, it has been used in a much wider range of applications. However, the second approach, based on the approximate recovery version of Fano's inequality in Theorem 2, has recently been shown to provide added flexibility in the reduction [35].

5.2.1 Reduction with Exact Recovery

As we discussed in Section 1, we seek to reduce the continuous problem to multiple hypothesis testing in such a way that successful minimax estimation implies success in the hypothesis test with high probability. To this end, we choose a *hard subset* $\theta_1, \ldots, \theta_M$, for which the elements are sufficiently well-separated so that the index $v \in \{1, \ldots, M\}$ can be identified from the estimate $\hat{\theta}$ (*cf.*, Figure 1). This is formalized in the proof of the following result.

Theorem 9. (Minimax bound via reduction to exact recovery) *Under the preceding minimax estimation setup, fix $\epsilon > 0$, and let $\{\theta_1, \ldots, \theta_M\}$ be a finite subset of Θ such that*

$$\rho(\theta_v, \theta_{v'}) \geq \epsilon, \quad \forall v, v' \in \{1, \ldots, M\}, v \neq v'. \tag{62}$$

Then, we have

$$\mathcal{M}_n(\Theta, \ell) \geq \Phi\Big(\frac{\epsilon}{2}\Big)\bigg(1 - \frac{I(V; \mathbf{Y}) + \log 2}{\log M}\bigg), \tag{63}$$

where V is uniform on $\{1, \dots, M\}$, and the mutual information is with respect to $V \to \theta_V \to \mathbf{Y}$. Moreover, in the special case $M = 2$, we have

$$\mathcal{M}_n(\Theta, \ell) \geq \Phi\Big(\frac{\epsilon}{2}\Big) H_2^{-1}\big(\log 2 - I(V; \mathbf{Y})\big), \tag{64}$$

where $H_2^{-1}(\cdot) \in [0, 0.5]$ is the inverse binary entropy function.

Proof. As illustrated in Figure 1, the idea is to reduce the estimation problem to a multiple hypothesis testing problem. As an initial step, we note from Markov's inequality that, for any $\epsilon_0 > 0$,

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta\big[\ell(\theta, \hat{\theta})\big] \geq \sup_{\theta \in \Theta} \Phi(\epsilon_0) \mathbb{P}_\theta[\ell(\theta, \hat{\theta}) \geq \Phi(\epsilon_0)] \tag{65}$$

$$= \Phi(\epsilon_0) \sup_{\theta \in \Theta} \mathbb{P}_\theta[\rho(\theta, \hat{\theta}) \geq \epsilon_0], \tag{66}$$

where (66) uses (60) and the assumption that $\Phi(\cdot)$ is increasing.

Suppose that a random index V is drawn uniformly from $\{1, \dots, M\}$, the samples $\mathbf{Y}$ are drawn from the distribution P_θ^n corresponding to $\theta = \theta_V$, and the estimator is applied to produce $\hat{\theta}$. Let $\hat{V}$ correspond to the closest θ_j according to the metric ρ, i.e., $\hat{V} = \arg\min_{v=1,\dots,M} \rho(\theta_v, \hat{\theta})$. Using the triangle inequality and the assumption (62), if $\rho(\theta_v, \hat{\theta}) < \frac{\epsilon}{2}$ then we must have $\hat{V} = v$; hence,

$$\mathbb{P}_v\left[\rho(\theta_v, \hat{\theta}) \geq \frac{\epsilon}{2}\right] \geq \mathbb{P}_v[\hat{V} \neq v], \tag{67}$$

where $\mathbb{P}_v$ is a shorthand for $\mathbb{P}_{\theta_v}$.

With the above tools in place, we proceed as follows:

$$\sup_{\theta \in \Theta} \mathbb{P}_\theta\left[\rho(\theta, \hat{\theta}) \geq \frac{\epsilon}{2}\right] \geq \max_{v=1,\dots,M} \mathbb{P}_v\left[\rho(\theta_v, \hat{\theta}) \geq \frac{\epsilon}{2}\right] \tag{68}$$

$$\geq \max_{v=1,\dots,M} \mathbb{P}_v[\hat{V} \neq v] \tag{69}$$

$$\geq \frac{1}{M} \sum_{v=1,\dots,M} \mathbb{P}_v[\hat{V} \neq v] \tag{70}$$

$$\geq 1 - \frac{I(V; \mathbf{Y}) + \log 2}{\log M}, \tag{71}$$

where (68) follows by maximizing over a smaller set, (69) follows from (67), (70) lower bounds the maximum by the average, and (71) follows from Fano's inequality (*cf.*, Theorem 1) and the fact that $I(V; \hat{V}) \leq I(V; \mathbf{Y})$ by the data processing inequality (*cf.*, Lemma 1).

The proof of (63) is concluded by substituting (71) into (66) with $\epsilon_0 = \frac{\epsilon}{2}$, and taking the infimum over all estimators $\hat{\theta}$. For $M = 2$, we obtain (64) in the same way upon replacing (71) by the version of Fano's inequality for $M = 2$ given in Remark 1. □

We return to this result in Section 5.3, where we introduce and compare some of the most widely-used approaches to choosing the set $\{\theta_1, \dots, \theta_M\}$ and bounding the mutual information.

5.2.2 Reduction with Approximate Recovery

The following generalization of Theorem 9, based on Fano's inequality with approximate recovery (*cf.*, Theorem 2), provides added flexibility in the reduction. An example comparing the two approaches will be given in

Section 6 for the sparse linear regression problem.

Theorem 10. (Minimax bound via reduction to approximate recovery) *Under the preceding minimax estimation setup, fix $\epsilon > 0$, $t \in \mathbb{R}$, a finite set $\mathcal{V}$ of cardinality M, and an arbitrary real-valued function $d(v, v')$ on $\mathcal{V} \times \mathcal{V}$, and let $\{\theta_v\}_{v \in \mathcal{V}}$ be a finite subset of Θ such that*

$$d(v, v') > t \implies \rho(\theta_v, \theta_{v'}) \geq \epsilon, \quad \forall v, v' \in \mathcal{V}. \tag{72}$$

Then we have for any $\epsilon \geq 0$ that

$$\mathcal{M}_n(\Theta, \ell) \geq \Phi\Big(\frac{\epsilon}{2}\Big)\bigg(1 - \frac{I(V; \mathbf{Y}) + \log 2}{\log \frac{M}{N_{\max}(t)}}\bigg), \tag{73}$$

where V is uniform on $\{1, \ldots, M\}$, the mutual information is with respect to $V \to \theta_V \to \mathbf{Y}$, and $N_{\max}(t) = \max_{v' \in \mathcal{V}} \sum_{v \in \mathcal{V}} \mathbb{1}\{d(v, v') \leq t\}$.

The proof is analogous to that of Theorem 9, and can be found in [35].

5.3 Local vs. Global Approaches

Here we highlight two distinct approaches to applying the reduction to exact recovery as per Theorem 9, termed the *local* and *global* approaches. We do not make such a distinction for the approximate recovery variant in Theorem 10, since we are not aware of a global approach being used previously for this variant.

Local approach. The most common approach to applying Theorem 9 is to construct a set $\{\theta_1, \ldots, \theta_M\}$ of elements that are "close" in KL divergence. Specifically, upper bounding the mutual information via Lemma 4 (with the vector $\mathbf{Y}$ playing the role of Y therein), one can weaken (63) as follows.

Corollary 1. (Local approach to minimax estimation) *Under the setup of Theorem 9 with a given set $\{\theta_1, \ldots, \theta_M\}$ satisfying (62), it holds for any auxiliary distribution $Q^n(\mathbf{y})$ that*

$$\mathcal{M}_n(\Theta, \ell) \geq \Phi\Big(\frac{\epsilon}{2}\Big)\bigg(1 - \frac{\min_{v=1,\ldots,M} D(P^n_{\theta_v} \| Q^n) + \log 2}{\log M}\bigg). \tag{74}$$

Moreover, the same bound holds true when $\min_v D(P^n_{\theta_v} \| Q^n)$ is replaced by any of $\frac{1}{M}\sum_v D(P^n_{\theta_v} \| Q)$, $\frac{1}{M^2}\sum_{v,v'} D(P^n_{\theta_v} \| P^n_{\theta_{v'}})$, or $\max_{v,v'} D(P^n_{\theta_v} \| P^n_{\theta_{v'}})$.

Attaining a good bound in (74) requires choosing $\{\theta_1, \ldots, \theta_M\}$ to trade off two competing objectives: (i) A larger value of M means that more hypotheses need to be distinguished; and (ii) A smaller value of $\min_v D(P^n_{\theta_v} \| Q^n)$ means that the hypotheses are more similar. Generally speaking, there is no single best approach to optimizing this trade-off, and the size and structure of the set can vary significantly from problem to problem. Moreover, the construction need not be explicit; one can instead use probabilistic arguments to prove the existence of a set satisfying the desired properties. Examples are given in Section 6. Naturally, an analog of Corollary 1 holds for $M = 2$ as per Theorem 9, and a counterpart for approximate recovery holds as per Theorem 10.

We briefly mention that Corollary 1 has interesting connections with the popular Assouad method from the statistics literature, as detailed in [47]. In addition, the counterpart of Corollary 1 with $M = 2$ is similarly related to an analogous technique known as Le Cam's method.

Global approach. An alternative approach to applying Theorem 9 is the *global* approach, which performs the following: (i) Construct a subset of Θ with as many elements as possible subject to the assumption (62);

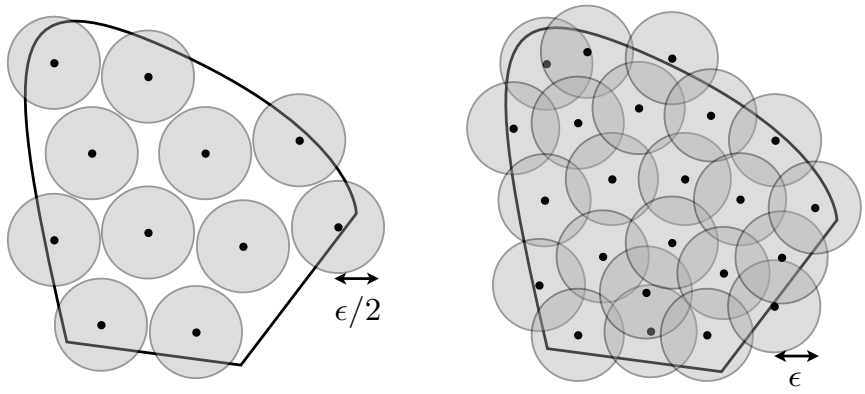

Figure 3: Examples of ϵ-packing (Left) and ϵ-covering (Right) sets in the case that ρ_0 is the Euclidean distance in $\mathbb{R}^2$. Since ρ_0 is a metric, a set of points is an ϵ-packing if and only if their corresponding $\frac{\epsilon}{2}$-balls do not intersect.

(ii) Construct a set that *covers* Θ, in the sense of Lemma 5, with as few elements as possible. The following definitions formalize the notions of forming "as many" and "as few" elements as possible. We write these in terms of a general real-valued function $\rho_0(\theta, \theta')$ that need not be a metric.

Definition 1. A set $\{\theta_1, \ldots, \theta_M\} \subseteq \Theta$ is said to be an ϵ_{p}-*packing set of* Θ *with respect to a measure* $\rho_0 : \Theta \times \Theta \to \mathbb{R}$ if $\rho_0(\theta_v, \theta_{v'}) \geq \epsilon_{\mathrm{p}}$ for all $v, v' \in \{1, \ldots, M\}$ with $v' \neq v$. The ϵ_{p}-*packing number* $M^*_{\rho_0}(\Theta, \epsilon_{\mathrm{p}})$ is defined to be the maximum cardinality of any ϵ_{p}-packing.

Definition 2. A set $\{\theta_1, \ldots, \theta_N\} \subseteq \Theta$ is said to be an ϵ_{c}-*covering set of* Θ *with respect to* $\rho_0 : \Theta \times \Theta \to \mathbb{R}$ if, for any $\theta \in \Theta$, there exists some $v \in \{1, \ldots, N\}$ such that $\rho_0(\theta, \theta_v) \leq \epsilon_{\mathrm{c}}$. The ϵ_{c}-*covering number* $N^*_{\rho_0}(\Theta, \epsilon_{\mathrm{c}})$ is defined to be the minimum cardinality of any ϵ_{c}-covering.

Observe that assumption (62) of Theorem 9 precisely states that $\{\theta_1, \ldots, \theta_M\}$ is an ϵ-packing set, though the result is often applied with M far smaller than the ϵ-packing number. The logarithm of the covering number is often referred to as the *metric entropy*.

The notions of packing and covering are illustrated in Figure 3. We do not explore the properties of packing and covering numbers in detail in this chapter; the interested reader is referred to [48, 49] for a more detailed treatment. We briefly state the following useful property, showing that the two definitions are closely related in the case that ρ_0 is a metric.

Lemma 7. (Packing vs. covering numbers) *If* ρ_0 *is a metric, then* $M^*_{\rho_0}(\Theta, 2\epsilon) \leq N^*_{\rho_0}(\Theta, \epsilon) \leq M^*_{\rho_0}(\Theta, \epsilon)$.

We now show how to use Theorem 9 to construct a lower bound on the minimax risk in terms of certain packing and covering numbers. For the packing number, we will directly consider the metric ρ used in Theorem 9. On the other hand, for the covering number, we consider the density $P^n_{\theta_v}(\mathbf{y})$ associated with each $\theta \in \Theta$, and use the associated KL divergence measure:

$$N^*_{\mathrm{KL},n}(\Theta, \epsilon) = N^*_{\rho^n_{\mathrm{KL}}}(\Theta, \epsilon), \quad \rho^n_{\mathrm{KL}}(\theta, \theta') = D(P^n_\theta \| P^n_{\theta'}). \tag{75}$$

Corollary 2. (Global approach to minimax estimation) *Under the minimax estimation setup of Section 5.1,*

we have for any $\epsilon_{\mathrm{p}} > 0$ *and* $\epsilon_{\mathrm{c},n} > 0$ *that*

$$\mathcal{M}_n(\Theta, \ell) \geq \Phi\Big(\frac{\epsilon_{\mathrm{p}}}{2}\Big)\bigg(1 - \frac{\log N^*_{\mathrm{KL},n}(\Theta, \epsilon_{\mathrm{c},n}) + \epsilon_{\mathrm{c},n} + \log 2}{\log M^*_{\rho}(\Theta, \epsilon_{\mathrm{p}})}\bigg). \tag{76}$$

In particular, if $P^n_\theta(\mathbf{y})$ *is the n-fold product of some single-measurement distribution* $P_\theta(y)$ *for each* $\theta \in \Theta$*, then we have for any* $\epsilon_{\mathrm{p}} > 0$ *and* $\epsilon_{\mathrm{c}} > 0$ *that*

$$\mathcal{M}_n(\Theta, \ell) \geq \Phi\Big(\frac{\epsilon_{\mathrm{p}}}{2}\Big)\bigg(1 - \frac{\log N^*_{\mathrm{KL}}(\Theta, \epsilon_{\mathrm{c}}) + n\epsilon_{\mathrm{c}} + \log 2}{\log M^*_{\rho}(\Theta, \epsilon_{\mathrm{p}})}\bigg), \tag{77}$$

where $N^*_{\mathrm{KL}}(\Theta, \epsilon) = N^*_{\rho_{\mathrm{KL}}}(\Theta, \epsilon)$ *with* $\rho_{\mathrm{KL}}(\theta, \theta') = D(P_\theta \| P_{\theta'})$.

Proof. Since Theorem 9 holds for any packing set, it holds for the maximal packing set. Moreover, using Lemma 5, we have $I(V; \mathbf{Y}) \leq \log N^*_{\mathrm{KL},n}(\Theta, \epsilon_{\mathrm{c},n}) + \epsilon_{\mathrm{c},n}$ in (63), since covering the entire space Θ is certainly enough to cover the elements in the packing set. Combining these, we obtain the first part of the corollary. The second part follows directly from the first part by choosing $\epsilon_{\mathrm{c},n} = n\epsilon_{\mathrm{c}}$ and noting that the KL divergence is additive for product distributions. □

Corollary 2 has been used as the starting point to derive minimax lower bounds for a wide range of problems [13]; see Section 6 for an example. It has been observed that the global approach is mainly useful for infinite-dimensional problems such as density estimation and non-parametric regression, with the local approach typically being superior for finite-dimensional problems such as vector or matrix estimation.

5.4 Beyond Estimation – Fano's Inequality for Optimization

While the minimax estimation framework captures a diverse range of problems of interest, there are also interesting problems that it does not capture. A notable example, which we consider in this section, is *stochastic optimization*. We provide a brief treatment, and refer the reader to [20] for further details and results.

We consider the following setup:

- We seek to minimize an unknown function $f : \mathcal{X} \to \mathbb{R}$ on some input domain $\mathcal{X}$, i.e., to find a point $x \in \mathcal{X}$ such that $f(x)$ is as low as possible.
- The algorithm proceeds in iterations: At the i-th iteration, a point $x_i \in \mathcal{X}$ is queried, and an *oracle* returns a sample y_i depending on the function, e.g., a noisy function value, a noisy gradient, or a tuple containing both. The selected point x_i can depend on the past queries and samples.
- After iteratively sampling n points, the optimization algorithm returns a final point $\hat{x}$, and the *loss* incurred is $\ell_f(\hat{x}) = f(\hat{x}) - \min_{x \in \mathcal{X}} f(x)$, i.e., the gap to the optimal function value.
- For a given class of functions $\mathcal{F}$, the *minimax risk* is given by

$$\mathcal{M}_n(\mathcal{F}) = \inf_{\hat{X}} \sup_{f \in \mathcal{F}} \mathbb{E}_f[\ell_f(\hat{X})], \tag{78}$$

  where the infimum is over all optimization algorithms that iteratively query the function n times and return a final point $\hat{x}$ as above, and $\mathbb{E}_f$ denotes expectation when the underlying function is f.

In the following, we let $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{Y} = (Y_1, \dots, Y_n)$ denote the queried locations and samples across the n rounds.

Theorem 11. (Minimax bound for noisy optimization) *Fix $\epsilon > 0$, and let $\{f_1, \dots, f_M\} \subseteq \mathcal{F}$ be a finite subset of $\mathcal{F}$ such that for each $x \in \mathcal{X}$, we have $\ell_{f_v}(x) \le \epsilon$ for at most one value of $v \in \{1, \dots, M\}$. Then we have*

$$\mathcal{M}_n(\mathcal{F}) \ge \epsilon \cdot \left(1 - \frac{I(V; \mathbf{X}, \mathbf{Y}) + \log 2}{\log M}\right), \tag{79}$$

where V is uniform on $\{1, \dots, M\}$, and the mutual information is with respect to $V \to f_V \to (\mathbf{X}, \mathbf{Y})$. Moreover, in the special case $M = 2$, we have

$$\mathcal{M}_n(\mathcal{F}) \ge \epsilon \cdot H_2^{-1}\big(\log 2 - I(V; \mathbf{X}, \mathbf{Y})\big), \tag{80}$$

where $H_2^{-1}(\cdot) \in [0, 0.5]$ is the inverse binary entropy function.

Proof. By Markov's inequality, we have

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f[\ell_f(\hat{X})] \ge \sup_{f \in \mathcal{F}} \epsilon \cdot \mathbb{P}_f[\ell_f(\hat{X}) \ge \epsilon]. \tag{81}$$

Suppose that a random index V is drawn uniformly from $\{1, \dots, M\}$, and the triplet $(\mathbf{X}, \mathbf{Y}, \hat{X})$ is generated by running the optimization algorithm on f_V. Given $\hat{X} = \hat{x}$, let $\hat{V}$ index the function among $\{f_1, \dots, f_M\}$ with the lowest corresponding value: $\hat{V} = \arg\min_{v=1,\dots,M} f_v(\hat{x})$.

By the assumption that any x satisfies $\ell_{f_v}(x) \le \epsilon$ for at most one of the M functions, we find that the condition $\ell_{f_v}(\hat{x}) \le \epsilon$ implies $\hat{V} = v$. Hence, we have

$$\mathbb{P}_v\big[\ell_{f_v}(\hat{X}) > \epsilon\big] \ge \mathbb{P}_{f_v}[\hat{V} \ne v]. \tag{82}$$

The remainder of the proof follows (68)–(71) in the proof of Theorem 9: We lower bound the minimax risk $\sup_{f \in \mathcal{F}} \mathbb{P}_f\big[\ell_f(\hat{X}) \ge \epsilon\big]$ by the average over V, and apply Fano's inequality (*cf.*, Theorem 1 and Remark 1) and the data processing inequality (*cf.*, third part of Lemma 3). □

Remark 7. Theorem 10 is based on reducing the optimization problem to a multiple hypothesis testing problem with exact recovery. One can derive an analogous result reducing to approximate recovery, but we are unaware of any works making use of such a result for optimization.

6 Applications – Continuous Settings

In this section, we present three applications of the tools introduced in Section 5: sparse linear regression, density estimation, and convex optimization. Similarly to the discrete case, our examples are chosen to permit a relatively simple analysis, while still effectively exemplifying the key concepts and tools.

6.1 Sparse Linear Regression

In this example, we extend the 1-sparse linear regression problem of Section 1.1 to the more general scenario of k-sparsity. The setup is described as follows:

- We wish to estimate a high-dimensional vector $\theta \in \mathbb{R}^p$ that is *k-sparse*: $\|\theta\|_0 \le k$, where $\|\theta\|_0$ is the number of non-zero entries in θ.
- The vector of n measurements is given by $\mathbf{Y} = \mathbf{X}\theta + \mathbf{Z}$, where $\mathbf{X} \in \mathbb{R}^{n\times p}$ is a known deterministic matrix, and $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$ is additive Gaussian noise.
- Given knowledge of $\mathbf{X}$ and $\mathbf{Y}$, an estimate $\hat{\theta}$ is formed, and the loss is given by the squared ℓ_2-error, $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$, corresponding to (60) with $\rho(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2$ and $\Phi(\cdot) = (\cdot)^2$. Overloading the general notation $\mathcal{M}_n(\Theta, \ell)$, we write the minimax risk as

$$\mathcal{M}_n(k, \mathbf{X}) = \inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^p \,:\, \|\theta_0\| \le k} \mathbb{E}_\theta[\|\theta - \hat{\theta}\|_2^2], \tag{83}$$

  where $\mathbb{E}_\theta$ denotes expectation when the underlying vector is θ.

6.1.1 Minimax Bound

The lower bound on the minimax risk is formally stated as follows. To simplify the analysis slightly, we state the result in an asymptotic form for the sparse regime $k = o(p)$; with only minor changes, one can attain a non-asymptotic variant attaining the same scaling laws for more general choices of k [35].

Theorem 12. (Sparse linear regression) *Under the preceding sparse linear regression problem with $k = o(p)$ and a fixed regression matrix $\mathbf{X}$, we have*

$$\mathcal{M}_n(k, \mathbf{X}) \ge \frac{\sigma^2 kp \log \frac{p}{k}}{8\|\mathbf{X}\|_F^2}(1 + o(1)) \tag{84}$$

as $p \to \infty$. In particular, under the constraint $\|\mathbf{X}\|_F^2 \le np\Gamma$ for some $\Gamma > 0$, achieving $\mathcal{M}_n(k, \mathbf{X}) \le \delta$ requires $n \ge \frac{\sigma^2 k \log \frac{p}{k}}{8\delta\Gamma}(1 + o(1))$.

Proof. We present a simple proof based on a reduction to approximate recovery (*cf.*, Theorem 10). In Section 6.1.2, we discuss an alternative proof based on a reduction to exact recovery (*cf.*, Theorem 9).

We define the set

$$\mathcal{V} = \big\{v \in \{-1, 0, 1\}^p \,:\, \|v\|_0 = k\big\}, \tag{85}$$

and to each $v \in \mathcal{V}$, we associate a vector $\theta_v = \epsilon' v$ for some $\epsilon' > 0$. Letting $d(v, v')$ denote the Hamming distance, we have the following properties:

- For $v, v' \in \mathcal{V}$, if $d(v, v') > t$, then $\|\theta_v - \theta_{v'}\|_2 > \epsilon'\sqrt{t}$;
- The cardinality of $\mathcal{V}$ is $|\mathcal{V}| = 2^k\binom{p}{k}$, yielding $\log|\mathcal{V}| \ge \log\binom{p}{k} \ge k \log\frac{p}{k}$;
- The quantity $N_{\max}(t)$ in Theorem 10 is the maximum possible number of $v' \in \mathcal{V}$ such that $d(v, v') \le t$ for a fixed v. Setting $t = \frac{k}{2}$, a simple counting argument gives $N_{\max}(t) \le \sum_{j=0}^{\lceil k/2\rceil} 2^j\binom{p}{j} \le \big(\lceil\frac{k}{2}\rceil + 1\big) \cdot 2^{\lceil k/2\rceil} \cdot \binom{p}{\lceil k/2\rceil}$, which simplifies to $\log N_{\max}(t) \le \big(\frac{k}{2}\log\frac{p}{k}\big)(1 + o(1))$ due to the assumption $k = o(p)$.

From these observations, applying Theorem 10 with $t = \frac{k}{2}$ and $\epsilon = \epsilon'\sqrt{\frac{k}{2}}$ yields

$$\mathcal{M}_n(k, \mathbf{X}) \ge \frac{k \cdot (\epsilon')^2}{2}\left(1 - \frac{I(V; \mathbf{Y}) + \log 2}{\big(\frac{k}{2}\log\frac{p}{k}\big)(1 + o(1))}\right). \tag{86}$$

Note that we do not condition on $\mathbf{X}$ in the mutual information, since we have assumed that $\mathbf{X}$ is deterministic.

To bound the mutual information, we first apply tensorization (*cf.*, first part of Lemma 2) to obtain $I(V;\mathbf{Y}) \le \sum_{i=1}^n I(V;Y_i)$, and then bound each $I(V;Y_i)$ using equation (24) in Lemma 4. We let Q_Y be the $\mathcal{N}(0,\sigma^2)$ density function, and we let $P_{v,i}$ denote the density function of $\mathcal{N}(X_i^T\theta_v,\sigma^2)$, where X_i is the transpose of the i-th row of $\mathbf{X}$. Since the KL divergence between the $\mathcal{N}(\mu_0,\sigma^2)$ and $\mathcal{N}(\mu_1,\sigma^2)$ density functions is $\frac{(\mu_1-\mu_0)^2}{2\sigma^2}$, we have $D(P_{v,i}\|Q_Y) = \frac{|X_i^T\theta_v|^2}{2\sigma^2}$. As a result, Lemma 4 yields $I(V;Y_i) \le \frac{1}{|\mathcal{V}|}\sum_v D(P_{v,i}\|Q_Y) = \frac{1}{2\sigma^2}\mathbb{E}\big[|X_i^T\theta_V|^2\big]$ for uniform V. Summing over i and recalling that $\theta_v = \epsilon' v$, we deduce that

$$I(V;\mathbf{Y}) \le \frac{(\epsilon')^2}{2\sigma^2}\mathbb{E}[\|\mathbf{X}V\|_2^2]. \tag{87}$$

From the choice of $\mathcal{V}$ in (85), we can easily compute $\mathrm{Cov}[V] = \frac{k}{p}\mathbf{I}_p$, which implies that $\mathbb{E}[\|\mathbf{X}V\|_2^2] = \frac{k}{p}\|\mathbf{X}\|_F^2$. Substitution into (87) yields $I(V;\mathbf{Y}) \le \frac{(\epsilon')^2}{2\sigma^2}\cdot\frac{k}{p}\|\mathbf{X}\|_F^2$, and we conclude from (86) that

$$\mathcal{M}_n(k,\mathbf{X}) \ge \frac{k\cdot(\epsilon')^2}{2}\left(1 - \frac{\frac{(\epsilon')^2}{2\sigma^2}\cdot\frac{k}{p}\|\mathbf{X}\|_F^2 + \log 2}{\big(\frac{k}{2}\log\frac{p}{k}\big)(1+o(1))}\right). \tag{88}$$

The proof is concluded by setting $(\epsilon')^2 = \frac{\sigma^2 p\log\frac{p}{k}}{2\|\mathbf{X}\|_F^2}$, which is chosen to make the bracketed term tend to $\frac{1}{2}$. □

Up to constant factors, the lower bound in Theorem 12 cannot be improved without additional knowledge of $\mathbf{X}$ beyond its Frobenius norm [5]. For instance, in the case that $\mathbf{X}$ has i.i.d. Gaussian entries, a matching upper bound holds with high probability under maximum-likelihood decoding.

6.1.2 Alternative Proof: Reduction with Exact Recovery

In contrast to the proof given above (adapted from [35]), the first known proof of Theorem 12 was based on packing with *exact* recovery (*cf.*, Theorem 9) [5]. For the sake of comparison, we briefly outline this alternative approach, which turns out to be more complicated.

The main step is to prove the existence of a set $\{\theta_1,\dots,\theta_M\}$ satisfying the following properties:

- The number of elements satisfies $M = \Omega\big(k\log\frac{p}{k}\big)$;
- Each element is k-sparse with non-zero entries equal to ± 1;
- The elements are well-separated in the sense that $\|\theta_v - \theta_{v'}\|_2^2 = \Omega(k)$ for $v \ne v'$;
- The empirical covariance matrix is close to a scaled identity matrix in the following sense: $\big\|\frac{1}{M}\sum_{v=1}^M \theta_v\theta_v^T - \frac{k}{p}\cdot\mathbf{I}_p\|_{2\to 2} = o\big(\frac{k}{p}\big)$, where $\|\cdot\|_{2\to 2}$ denotes the ℓ_2/ℓ_2-operator norm, i.e., the largest singular value.

Once this is established, the proof proceeds along the same lines as the proof we gave above, scaling the vectors down by some $\epsilon' > 0$ and using Theorem 9 in place of Theorem 10.

The existence of the packing set is proved via a probabilistic argument: If one generates $\Omega\big(k\log\frac{p}{k}\big)$ uniformly random k-sparse sequences with non-zero entries equaling ± 1, then these will satisfy the remaining two properties with positive probability. While it is straightforward to establish the condition of being well-separated, the proof of the condition on the empirical covariance matrix requires a careful application of the non-elementary matrix Bernstein inequality.

Overall, while the two approaches yield the same result up to constant factors in this example, the approach based on approximate recovery is entirely elementary and avoids the preceding difficulties.

6.2 Density Estimation

In this subsection, we consider the problem of estimating an entire probability density function given samples from its distribution, commonly known as *density estimation*. We consider a non-parametric view, meaning that the density does not take any specific parametric form. As a result, the problem is inherently infinite-dimensional, and lends itself to the global packing and covering approach introduced in Section 5.3.

While many classes of density functions have been considered in the literature [13], we focus our attention on a specific setting for clarity of exposition:

- The density function f that we seek to estimate is defined on the domain $[0,1]$, i.e., $f(y) \geq 0$ for all $y \in [0,1]$, and $\int_0^1 f(y)dy = 1$.
- We assume that f satisfies the following conditions:

$$f(y) \geq \eta, \forall y \in [0,1], \qquad \|f\|_{\mathrm{TV}} \leq \Gamma \tag{89}$$

  for some $\eta \in (0,1)$ and $\Gamma > 0$, where the *total variation* (TV) norm is defined as $\|f\|_{\mathrm{TV}} = \sup_L \sup_{0\leq x_1 \leq \ldots \leq x_L \leq 1} \sum_{l=2}^{L} \big(f(x_l) - f(x_{l-1})\big)$. The set of all density functions satisfying these constraints is denoted by $\mathcal{F}_{\eta,\Gamma}$.

- Given n independent samples $\mathbf{Y} = (Y_1, \ldots, Y_n)$ from f, an estimate $\hat{f}$ is formed, and the loss is given by $\ell(f, \hat{f}) = \|f - \hat{f}\|_2^2 = \int_0^1 (f(x) - \hat{f}(x))^2 dx$. Hence, the minimax risk is given by

$$\mathcal{M}_n(\eta, \Gamma) = \inf_{\hat{f}} \sup_{f \in \mathcal{F}_{\eta,\Gamma}} \mathbb{E}_f\big[\|f - \hat{f}\|_2^2\big], \tag{90}$$

  where $\mathbb{E}_f$ denotes expectation when the underlying density is f.

6.2.1 Minimax Bound

The minimax lower bound is given is follows.

Theorem 13. (Density estimation) *Consider the preceding density estimation setup with some $\eta \in (0,1)$ and $\Gamma > 0$ not depending on n. There exists a constant $c > 0$ (depending on η and Γ) such that in order to achieve $\mathcal{M}_n(\eta, \Gamma) \leq \delta$, it is necessary that*

$$n \geq c \cdot \Big(\frac{1}{\delta}\Big)^{3/2} \tag{91}$$

when δ is sufficiently small. In other words, $\mathcal{M}_n(\eta, \Gamma) = \Omega\big(n^{-2/3}\big)$.

Proof. We specialize the general analysis of [13] to the class $\mathcal{F}_{\eta,\Gamma}$. Recalling the packing and covering numbers from Definitions 1 and 2, we adopt the shorthand notation $M_2^*(\epsilon_{\mathrm{p}}) = M_\rho^*(\mathcal{F}_{\eta,\Gamma}, \epsilon_{\mathrm{p}})$ with $\rho(f, f') = \|f - f'\|_2$, and similarly $N_2^*(\epsilon_{\mathrm{c}}) = N_\rho^*(\mathcal{F}_{\eta,\Gamma}, \epsilon_{\mathrm{c}})$. We first show that N_{KL}^* (*cf.*, Corollary 2) can be upper bounded in terms of M_2^*, which will lead to a minimax lower bound that depends only on the packing number M_2^*. For $f_1, f_2 \in \mathcal{F}_{\eta,\Gamma}$, we have

$$D(f_1 \| f_2) \leq \int_0^1 \frac{(f_1(x) - f_2(x))^2}{f_2(x)} dx \tag{92}$$

$$\leq \frac{1}{\eta} \int_0^1 (f_1(x) - f_2(x))^2 dx \tag{93}$$

$$= \frac{1}{\eta} \|f_1 - f_2\|_2^2, \tag{94}$$

where (92) follows since the KL divergence is upper bounded by the χ^2-divergence (*cf.*, Lemma 6), and (93) follows from the assumption that the density is lower bounded by η. From the definition of N^*_{KL} in Corollary 2, we deduce the following for any $\epsilon_{\mathrm{c}} > 0$:

$$N^*_{\mathrm{KL}}(\epsilon_{\mathrm{c}}) \leq N^*_2(\sqrt{\eta\epsilon_{\mathrm{c}}}) \leq M^*_2(\sqrt{\eta\epsilon_{\mathrm{c}}}), \tag{95}$$

where the first inequality holds because any $\sqrt{\eta\epsilon_{\mathrm{c}}}$-covering the the ℓ_2-norm is also a ϵ_{c}-covering in the KL divergence due to (94), and the second inequality follows from Lemma 7.

Combining (95) with Corollary 2 and the choice $\Phi(\cdot) = (\cdot)^2$ gives

$$\mathcal{M}_n(\eta, \Gamma) \geq \left(\frac{\epsilon_{\mathrm{p}}}{2}\right)^2 \left(1 - \frac{\log M^*_2(\sqrt{\eta\epsilon_{\mathrm{c}}}) + n\epsilon_{\mathrm{c}} + \log 2}{\log M^*_2(\epsilon_{\mathrm{p}})}\right). \tag{96}$$

We now apply the following bounds on the packing number of $\mathcal{F}_{\eta,\Gamma}$, which we state from [13] without proof:

$$\underline{c} \cdot \epsilon^{-1} \leq \log M^*_2(\epsilon) \leq \overline{c} \cdot \epsilon^{-1}, \tag{97}$$

for some constants $\underline{c}, \overline{c} > 0$ and sufficiently small $\epsilon > 0$. It follows that

$$\mathcal{M}_n(\eta, \Gamma) \geq \left(\frac{\epsilon_{\mathrm{p}}}{2}\right)^2 \left(1 - \frac{\overline{c} \cdot (\eta\epsilon_{\mathrm{c}})^{-1/2} + n\epsilon_{\mathrm{c}} + \log 2}{\underline{c} \cdot \epsilon_{\mathrm{p}}^{-1}}\right). \tag{98}$$

The remainder of the proof amounts to choosing ϵ_{p} and ϵ_{c} to balance the terms appearing in this expression.

First, choosing ϵ_{c} to equate the terms $\overline{c} \cdot (\eta\epsilon_{\mathrm{c}})^{-1/2}$ and $n\epsilon_{\mathrm{c}}$ leads to $\epsilon_{\mathrm{c}} = \left(\frac{c'}{n}\right)^{2/3}$ with $c' = \overline{c}\eta^{-1/2}$, yielding $\frac{\overline{c}\cdot(\eta\epsilon_{\mathrm{c}})^{-1/2} + n\epsilon_{\mathrm{c}} + \log 2}{\underline{c}\cdot\epsilon_{\mathrm{p}}^{-1}} = \frac{2n\left(\frac{c'}{n}\right)^{2/3} + \log 2}{\overline{c}\cdot\epsilon_{\mathrm{p}}^{-1}}$. Next, choosing ϵ_{p} to make this fraction equal to $\frac{1}{2}$ yields $\epsilon_{\mathrm{p}}^{-1} = \frac{2}{\overline{c}}\big(2(c')^{2/3}n^{1/3} + \log 2\big)$, which means that $\epsilon_{\mathrm{p}} \geq c'' \cdot n^{-1/3}$ for suitable $c'' > 0$ and sufficiently large n. Finally, since we made the fraction equal to $\frac{1}{2}$, (98) yields $\mathcal{M}_n(\eta, \Gamma) \geq \frac{\epsilon_{\mathrm{p}}^2}{8} \geq \frac{(c'')^2 n^{-2/3}}{8}$. Setting $\mathcal{M}_n(\eta, \Gamma) = \delta$ and solving for n yields the desired result. □

The scaling given in Theorem 13 cannot be improved; a matching upper bound is given in [13], and can be achieved even when $\eta = 0$.

6.3 Convex Optimization

In our final example, we consider the optimization setting introduced in Section 5.4. We provide an example that is rather simple, yet has interesting features not present in the previous examples: (i) an example departing from estimation; (ii) a continuous example with adaptivity; and (iii) a case where Fano's inequality with $|\mathcal{V}| = 2$ is used.

We consider the following special case of the general setup of Section 5.4:

- We let $\mathcal{F}$ be the set of differentiable and *strongly convex* functions on $\mathcal{X} = [0, 1]$, with strong convexity parameter equal to one:

$$\mathcal{F}_{\mathrm{scv}} = \left\{ f \,:\, f \text{ is differentiable} \cap\, f(x) - \frac{1}{2}x^2 \text{ is convex} \right\}. \tag{99}$$

  The analysis that we present can easily be extended to functions on an arbitrary closed interval with an arbitrary strong convexity parameter.

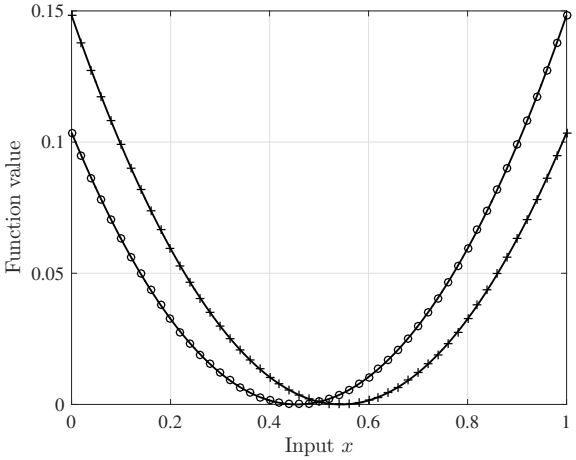

Figure 4: Construction of two functions in $\mathcal{F}_{\text{scv}}$ that are difficult to distinguish, and such that any point $x \in [0,1]$ can be ϵ-optimal for only one of the two functions.

- When we query a point $x \in \mathcal{X}$, we observe a noisy sample of the function value and its gradient:

$$Y = (f(x) + Z, f'(x) + Z'), \tag{100}$$

  where Z and Z' are independent $\mathcal{N}(0, \sigma^2)$ random variables, for some $\sigma^2 > 0$. This is commonly referred to as the *noisy first-order oracle*.

6.3.1 Minimax Bound

The following theorem lower bounds the number of queries required to achieve δ-optimality. The proof is taken from [20] with only minor modifications.

Theorem 14. (Stochastic optimization of strongly convex functions) *Under the preceding convex optimization setting with noisy first-order oracle information, in order to achieve $\mathcal{M}_n(\mathcal{F}_{\text{scv}}) \leq \delta$, it is necessary that*

$$n \geq \frac{\sigma^2 \log 2}{40\delta} \tag{101}$$

when δ is sufficiently small.

Proof. We construct a set of two functions satisfying the assumptions of Theorem 11. Specifically, we fix (ϵ, ϵ') such that $0 < \epsilon < \epsilon' < \frac{1}{8}$, define $x_1^* = \frac{1}{2} - \sqrt{2\epsilon'}$ and $x_2^* = \frac{1}{2} + \sqrt{2\epsilon'}$, and set

$$f_v(x) = \frac{1}{2}(x - x_v^*)^2, \quad v = 1, 2. \tag{102}$$

These functions are illustrated in Figure 4.

Since $\epsilon' \in \left(0, \frac{1}{8}\right)$, both x_1^* and x_2^* lie in $(0,1)$, and hence $\min_{x\in[0,1]} f_1(x) = \min_{x\in[0,1]} f_2(x) = 0$. Moreover, a direct evaluation reveals that $f_1(x) + f_2(x) = \left(x - \frac{1}{2}\right)^2 + 2\epsilon' > 2\epsilon$, which implies that any ϵ-optimal point for one function cannot be ϵ-optimal for the other function. This is the condition needed to apply Theorem 11, yielding from (80) that

$$\mathcal{M}_n(\mathcal{F}_{\text{scv}}) \geq \epsilon \cdot H_2^{-1}(\log 2 - I(V; \mathbf{X}, \mathbf{Y})). \tag{103}$$

To bound the mutual information, we first apply tensorization (*cf.*, first part of Lemma 3) to obtain

$I(V;\mathbf{X},\mathbf{Y}) \le \sum_{i=1}^n I(V;Y_i|X_i)$. We proceed by bounding $I(V;Y_i|X_i)$ for any given i. Fix $x \in [0,1]$, let P_{Y_x} and $P_{Y'_x}$ be the density functions of the noisy samples of $f_1(x)$ and $f'_1(x)$, and let Q_{Y_x} and $Q_{Y'_x}$ be defined similarly for $f_0(x) = \frac{1}{2}\big(x - \frac{1}{2}\big)^2$. We have

$$D(P_{Y_x} \times P_{Y'_x} \| Q_{Y_x} \times Q_{Y'_x}) = D(P_{Y_x}\|Q_{Y_x}) + D(P_{Y'_x}\|Q_{Y'_x}) \tag{104}$$

$$= \frac{(f_1(x) - f_0(x))^2}{2\sigma^2} + \frac{(f'_1(x) - f'_0(x))^2}{2\sigma^2}, \tag{105}$$

where (104) holds since the KL divergence is additive for product distributions, and (105) uses the fact that the divergence between the $\mathcal{N}(\mu_0,\sigma^2)$ and $\mathcal{N}(\mu_1,\sigma^2)$ density functions is $\frac{(\mu_1-\mu_0)^2}{2\sigma^2}$.

Recalling that $f_1(x) = \frac{1}{2}\big(x - \frac{1}{2} + \sqrt{2\epsilon'}\big)^2$ and $f_0(x) = \frac{1}{2}\big(x - \frac{1}{2}\big)^2$, we have

$$(f_1(x) - f_0(x))^2 = \frac{1}{4}\bigg(2\epsilon' + 2\Big(x - \frac{1}{2}\Big)\sqrt{2\epsilon'}\bigg)^2 \le \Big(\epsilon' + \sqrt{\frac{\epsilon'}{2}}\Big)^2 \le 2\epsilon', \tag{106}$$

where the first inequality uses the fact that $x \in [0,1]$, and the second inequality follows since $\epsilon' < \frac{1}{8}$ and hence $\epsilon' = \sqrt{\epsilon'} \cdot \sqrt{\epsilon'} \le \sqrt{\frac{\epsilon'}{8}}$ (note that $\big(\frac{1}{\sqrt{8}} + \frac{1}{\sqrt{2}}\big)^2 \le 2$). Moreover, taking the derivatives of f_0 and f_1 gives $(f'_1(x) - f'_0(x))^2 = 2\epsilon'$, and substitution into (105) yields $D(P_{Y_x} \times P_{Y'_x} \| Q_{Y_x} \times Q_{Y'_x}) \le \frac{2\epsilon'}{\sigma^2}$.

The preceding analysis applies in a near-identical manner when f_2 is used in place of f_1, and yields the same KL divergence bound when $(P_{Y_x}, P_{Y'_x})$ is defined with respect to f_2. As a result, for any $x \in [0,1]$, we obtain from (25) in Lemma 4 that $I(V;Y_i|X_i = x) \le \frac{2\epsilon'}{\sigma^2}$. Averaging over X, we obtain $I(V;Y_i|X_i) \le \frac{2\epsilon'}{\sigma^2}$, and substitution into the above-established bound $I(V;\mathbf{X},\mathbf{Y}) \le \sum_{i=1}^n I(V;Y_i|X_i)$ yields $I(V;\mathbf{X},\mathbf{Y}) \le \frac{2n\epsilon'}{\sigma^2}$. Hence, (103) yields

$$\mathcal{M}_n(\mathcal{F}_{\mathrm{scv}}) \ge \epsilon \cdot H_2^{-1}\bigg(\log 2 - \frac{2n\epsilon'}{\sigma^2}\bigg). \tag{107}$$

Now observe that if $n \le \frac{\sigma^2 \log 2}{4\epsilon'}$ then the argument to $H_2^{-1}(\cdot)$ is at least $\frac{\log 2}{2}$. It is easy to verify that $H_2^{-1}\big(\frac{\log 2}{2}\big) > \frac{1}{10}$, from which it follows that $\mathcal{M}_n(\mathcal{F}_{\mathrm{scv}}) > \frac{\epsilon}{10}$. Setting $\epsilon = 10\delta$ and noting that ϵ' can be chosen arbitrarily close to ϵ, we conclude that the required number of samples $\frac{\sigma^2 \log 2}{4\epsilon'}$ recovers (101). □

Theorem 14 provides tight scaling laws, since stochastic gradient descent is known to achieve δ-optimality for strongly convex functions using $O\big(\frac{\sigma^2}{\delta}\big)$ queries. Analogous results for the multi-dimensional setting can be found in [20].

7 Discussion

7.1 Limitations of Fano's Inequality

While Fano's inequality is a highly versatile method with successes in a wide range of statistical applications (*cf.*, Table 1), it is worth pointing out some of its main limitations. We briefly mention some alternative methods below, as well as discussing some suitable generalizations of Fano's inequality in Section 7.2.

Non-asymptotic weakness. Even in scenarios where Fano's inequality provides converse bounds with the correct asymptotics including constants, these bounds can be inferior to alternative methods in the non-asymptotic sense [50,51]. Related to this issue is the distinction between the *weak converse* and *strong converse*: We have seen that Fano's inequality typically provides necessary conditions of the form $n \ge n^*(1 - \delta - o(1))$ for achieving $P_{\mathrm{e}} \le \delta$, in contrast with strong converse results of the form $n \ge n^*(1 - o(1))$ for *any* $\delta \in (0,1)$.

Alternative techniques addressing these limitations are discussed in the context of communication in [50], and in the context of statistical estimation in [52, 53].

Difficulties in adaptive settings. While we have provided examples where Fano's inequality provides tight bounds in adaptive settings, there are several applications where alternative methods have proved to be more suitable. One reason for this is that the conditional mutual information terms $I(V;Y_i|X_i)$ (*cf.*, Lemma 3) often involve complicated conditional distributions that are difficult to analyze. We refer the reader to [54–56] for examples in which alternative techniques proved to be more suitable for adaptive settings.

Restriction to KL divergence. When applying Fano's inequality, one invariably needs to bound a mutual information term, which is an instance of the KL divergence. While the KL divergence satisfies a number of convenient properties that can help in this process, it is sometimes the case that other divergence measures are more convenient to work with, or can be used to derive tighter results. Generalizations of Fano's inequality have been proposed specifically for this purpose, as we discuss in the following subsection.

7.2 Generalizations of Fano's Inequality

Several variations and generalizations of Fano's inequality have been proposed in the literature [57–62]. Most of these are not derived based on the most well-known proof of Theorem 1, but are instead based on an alternative proof via the data processing inequality for KL divergence: For any event E, one has

$$I(V;\hat{V}) = D(P_{V\hat{V}} \| P_V \times P_{\hat{V}}) \geq D_2\big(P_{V\hat{V}}[E] \,\|\, (P_V \times P_{\hat{V}})[E]\big), \tag{108}$$

where $D_2(p\|q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$ is the binary KL divergence function. Observe that if V is uniform and E is the event that $V \neq \hat{V}$, then we have $P_{V\hat{V}}[E] = P_{\mathrm{e}}$ and $(P_V \times P_{\hat{V}})[E] = 1 - \frac{1}{|\mathcal{V}|}$, and Fano's inequality (*cf.*, Theorem 1) follows by substituting the definition of $D_2(\cdot\|\cdot)$ in (108) and re-arranging. This proof lends itself to interesting generalizations, including the following.

Continuum version. Consider a continuous random variable V taking values on $\mathcal{V} \subseteq \mathbb{R}^p$ for some $p \geq 1$, and an error probability of the form $P_{\mathrm{e}}(t) = \mathbb{P}\big[d(V,\hat{V}) > t\big]$ for some real-valued function d on $\mathbb{R}^p \times \mathbb{R}^p$. This is the same formula as (11), which we previously introduced for the discrete setting. Defining the "ball" $\mathbb{B}_d(\hat{v},t) = \{v \in \mathbb{R}^p : d(v,\hat{v}) \leq t\}$ centered at $\hat{v}$, (108) leads to the following for V uniform on $\mathcal{V}$:

$$P_{\mathrm{e}}(t) \geq 1 - \frac{I(V;\hat{V}) + \log 2}{\log \frac{\mathrm{Vol}(\mathcal{V})}{\sup_{\hat{v}\in\mathbb{R}^p} \mathrm{Vol}(\mathcal{V} \cap \mathbb{B}_d(\hat{v},t))}}, \tag{109}$$

where $\mathrm{Vol}(\cdot)$ denotes the volume of a set. This result provides a continuous counterpart to the final part of Theorem 2, in which the cardinality ratio is replaced by a volume ratio. We refer the reader to [35] for example applications, and to [62] for the simple proof outlined above.

Beyond KL divergence. The key step (108) extends immediately to other measures that satisfy the data processing inequality. A useful class of such measures is the class of *f-divergences*: $D_f(P\|Q) = \mathbb{E}_Q\big[f\big(\frac{P(\mathbf{Y})}{Q(\mathbf{Y})}\big)\big]$ for some convex f satisfying $f(1) = 0$. Special cases include KL divergence ($f(z) = z\log z$), total variation ($f(z) = \frac{1}{2}|z-1|$), squared Hellinger distance ($f(z) = (\sqrt{z}-1)^2$), and χ^2-divergence ($f(z) = (z-1)^2$). It was shown in [60] that alternative choices beyond the KL divergence can provide improved bounds in some cases. Generalizations of Fano's inequality beyond f-divergences can be found in [61].

Non-uniform priors. The first form of Fano's inequality in Theorem 1 does not require V to be uniform. However, in highly non-uniform cases where $H(V) \ll \log|\mathcal{V}|$, the term $P_{\mathrm{e}}\log(|\mathcal{V}|-1)$ may be too large for the bound to be useful. In such cases, it is often useful to use different Fano-like bounds based on the alternative

proof above. In particular, the step (108) makes no use of uniformity, and continues to hold even in the non-uniform case. In [57], this bound was further weakened to provide simpler lower bounds for non-uniform settings with discrete alphabets. Fano-type lower bounds in *continuous* Bayesian settings with non-uniform priors arose more recently, and are typically more technically challenging; the interested reader is referred to [18, 63].

A Appendix

Here we provide the omitted proofs from the main body. Throughout the proofs, the random variables V and $\hat{V}$ are assumed to be discrete, whereas the other random variables involved, including the inputs $\mathbf{X} = (X_1, \ldots, X_n)$ and samples $\mathbf{Y} = (Y_1, \ldots, Y_n)$, may be continuous. In such cases, entropy quantities such as $H(Y_i)$ should be interpreted as being the *differential entropy* [34, Ch. 8], and probability functions such as $P_Y(y)$ should be interpreted as being a probability density function (PDF).

A.1 Preliminary Information-Theoretic Results

The following lemma states some useful results from information theory. The proofs can be found in standard references such as [34].

Lemma 8. (Standard information-theoretic results) *We have the following:*

- (Chain rule for entropy) $H(Y_1, \ldots, Y_n) = \sum_{i=1}^{n} H(Y_i|Y_1, \ldots, Y_{i-1})$.
- (Chain rule for mutual information) $I(X; Y_1, \ldots, Y_n) = \sum_{i=1}^{n} I(X; Y_i|Y_1, \ldots, Y_{i-1})$.
- (Sub-additivity of entropy) $H(Y_1, \ldots, Y_n) \leq \sum_{i=1}^{n} H(Y_i)$.
- (Conditioning reduces entropy) $H(Y|X) \leq H(Y)$.
- (Information-preserving transform) *If Y depends on X only through $f(X)$, then $H(Y|X, f(X)) = H(Y|f(X))$, and $I(X;Y) = I(f(X);Y)$.*
- (Capacity of binary symmetric channel) *If X, Y are binary with $Y = X \oplus Z$ for $Z \sim$ Bernoulli(ϵ) (where $\oplus$ denotes modulo-2 addition), then $I(X;Y) \leq \log 2 - H_2(\epsilon)$.*
- (Divergence between independent pairs) $D(P_X \times P_Y \| Q_X \times Q_Y) = D(P_X \| Q_X) + D(P_Y \| Q_Y)$.
- (Divergence between equal-variance univariate Gaussians) *For $X \sim \mathcal{N}(\mu_1, \sigma^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma^2)$, it holds that $D(P_X \| P_Y) = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2}$.*

We will make use of these results without necessarily referencing the lemma.

A.2 Proof of Theorem 1 (Fano's Inequality)

Defining the error indicator random variable $E = \mathbb{1}\{V \neq \hat{V}\}$, we have

$$\begin{aligned} H(V|\hat{V}) &= H(V, E|\hat{V}) && (110)\\ &= H(E|\hat{V}) + H(V|\hat{V}, E) && (111)\\ &\leq H(E) + H(V|\hat{V}, E) && (112)\\ &= H_2(P_\mathrm{e}) + P_\mathrm{e} H(V|\hat{V}, E=1) + (1-P_\mathrm{e}) H(V|\hat{V}, E=0) && (113)\\ &= H_2(P_\mathrm{e}) + P_\mathrm{e} \log\big(|\mathcal{V}| - 1\big), && (114) \end{aligned}$$

where (110) holds since E is a deterministic function of $(V, \hat{V})$, (111) follows from the chain rule, (112) holds since conditioning reduces entropy, (113) uses $H(E) = H_2(P_\mathrm{e})$, and (114) follows since V has no uncertainty given $\hat{V}$ when $E = 0$, and takes one of $|\mathcal{V}| - 1$ values given $\hat{V}$ when $E = 1$.

In case that V is uniform, we obtain (7) by upper bounding $|\mathcal{V}| - 1 \leq |\mathcal{V}|$ and $H_2(P_\mathrm{e}) \leq \log 2$ in (6), subtracting $H(V) = \log|\mathcal{V}|$ on both sides, and taking the negative on both sides.

A.3 Proof of Theorem 2 (Fano's Inequality with Approximate Recovery)

Define the error event $E_t = \{d(V, \hat{V}) > t\}$. Following the steps (110)–(113) with E_t in place of E, we obtain

$$\begin{aligned} H(V|\hat{V}) &\leq H_2(P_\mathrm{e}(t)) + P_\mathrm{e}(t) H(V|\hat{V}, E_t = 1) + (1 - P_\mathrm{e}(t)) H(V|\hat{V}, E_t = 0) && (115)\\ &\leq H_2(P_\mathrm{e}(t)) + P_\mathrm{e}(t) \log\big(|\mathcal{V}| - N_{\min}(t)\big) + (1 - P_\mathrm{e}(t)) \log N_{\max}(t) && (116)\\ &= H_2(P_\mathrm{e}(t)) + P_\mathrm{e}(t) \log \frac{|\mathcal{V}| - N_{\min}(t)}{N_{\max}(t)} + \log N_{\max}(t), && (117) \end{aligned}$$

where (116) follows since when $\hat{V}$ is given and $E_t = 0$, V takes one of at most $N_{\max}(t)$ values, whereas if $\hat{V}$ is given and $E_t = 1$, V takes one of at most $|\mathcal{V}| - N_{\min}(t)$ values. We have thus proved (14).

In case that V is uniform, we obtain (15) by upper bounding $|\mathcal{V}| - N_{\min}(t) \leq |\mathcal{V}|$ and $H_2(P_\mathrm{e}(t)) \leq \log 2$ in (14), subtracting $H(V) = \log|\mathcal{V}|$ on both sides, and taking the negative on both sides.

A.4 Proof of Lemma 1 (Data Processing Inequality)

We focus on the first part, since the second and third parts follow as special cases. We have

$$\begin{aligned} I(V; \hat{V}) &= H(V) - H(V|\hat{V}) && (118)\\ &\leq H(V) - H(V|\hat{V}, \mathbf{Y}) && (119)\\ &= H(V) - H(V|\mathbf{Y}) && (120)\\ &= I(V; \mathbf{Y}), && (121) \end{aligned}$$

where (119) follows since conditioning reduces entropy, and (120) holds because V and $\hat{V}$ are conditionally independent given $\mathbf{Y}$.

A.5 Proof of Lemma 2 (Tensorization)

We start with the second claim, since the first claim then follows by letting each X_i deterministically equal an arbitrary fixed value (e.g., zero). To prove the second claim, we write

$$\begin{aligned} I(V;\mathbf{Y}|\mathbf{X}) &= H(\mathbf{Y}|\mathbf{X}) - H(\mathbf{Y}|V,\mathbf{X}) && (122)\\ &\le \sum_{i=1}^{n} H(Y_i|X_i) - H(\mathbf{Y}|V,\mathbf{X}) && (123)\\ &= \sum_{i=1}^{n} \big(H(Y_i|X_i) - H(Y_i|V,\mathbf{X})\big) && (124)\\ &= \sum_{i=1}^{n} \big(H(Y_i|X_i) - H(Y_i|V,X_i)\big) && (125)\\ &= \sum_{i=1}^{n} I(V;Y_i|X_i), && (126) \end{aligned}$$

where (123) follows from the sub-additivity of entropy and the fact that conditioning reduces entropy, (124) follows from the conditional independence of the Y_i given $(V,\mathbf{X})$, and (125) follows from the assumption that Y_i depends on $(V,\mathbf{X})$ only on through (V,X_i).

The third claim follows from the second claim by writing

$$I(V;Y_i|X_i) \le I(V,X_i;Y_i) = I(U_i;Y_i) \tag{127}$$

by the assumption that Y_i depends on (V,X_i) only through U_i.

A.6 Proof of Lemma 3 (Tensorization with Adaptivity)

We have the following:

$$\begin{aligned} I(V;\mathbf{X},\mathbf{Y}) &= \sum_{i=1}^{n} I(X_i,Y_i;V \mid X_1^{i-1},Y_1^{i-1}) && (128)\\ &= \sum_{i=1}^{n} I(Y_i;V \mid X_1^{i-1},Y_1^{i-1},X_i) && (129)\\ &= \sum_{i=1}^{n} \Big(H(Y_i \mid X_1^{i-1},Y_1^{i-1},X_i) - H(Y_i \mid X_1^{i-1},Y_1^{i-1},X_i,V)\Big) && (130)\\ &= \sum_{i=1}^{n} \Big(H(Y_i \mid X_1^{i-1},Y_1^{i-1},X_i) - H(Y_i \mid V,X_i)\Big) && (131)\\ &\le \sum_{i=1}^{n} \Big(H(Y_i \mid X_i) - H(Y_i|V,X_i)\Big) && (132)\\ &= \sum_{i=1}^{n} I(V;Y_i|X_i), && (133) \end{aligned}$$

where (128) follows from the chain rule, (129) follows since X_i is a function of (X_1^{i-1},Y_1^{i-1}), (131) follows since Y_i is conditionally independent of (X_1^{i-1},Y_1^{i-1}) given (V,X_i), and (132) follows since conditioning reduces entropy. This completes the proof of the first part.

To prove the second part, we note that

$$I(V;Y_i|X_i) \le I(V,X_i;Y_i) = I(U_i;Y_i) \tag{134}$$

by the assumption that Y_i depends on (X_i, V) only through U_i.

A.7 Proof of Lemma 5 (Covering-Based Mutual Information Bound)

Applying (25) in Lemma 4 with the choice $Q_Y(y) = \frac{1}{N}\sum_{j=1}^N Q_j(y)$, and letting $\mathbb{E}_v$ denote expectation with respect to $P_{Y|V}(\cdot\,|v)$, we have

$$\begin{aligned}
I(V;Y) &\le \max_v D\bigg(P_{Y|V}(\cdot\,|\,v)\,\bigg\|\,\frac{1}{N}\sum_{j=1}^N Q_j\bigg) && (135)\\
&= \max_v \mathbb{E}_v\bigg[\log\frac{P_{Y|V}(Y\,|\,v)}{\frac{1}{N}\sum_{j=1}^N Q_j(Y)}\bigg] && (136)\\
&\le \max_v \mathbb{E}_v\bigg[\log\frac{P_{Y|V}(Y\,|\,v)}{\frac{1}{N}Q_{j^*(v)}(Y)}\bigg] && (137)\\
&= \log N + \max_v D\big(P_{Y|V}(\cdot\,|\,v)\,\big\|\,Q_{j^*(v)}\big) && (138)\\
&\le \log N + \epsilon, && (139)
\end{aligned}$$

where (136) applies the definition of KL divergence, (137) lower bounds the summation by the single term $j^*(v)$ achieving the minimum in (30), and (139) applies the upper bound in (30).

A.8 Omitted Details in Discrete Examples with Approximate Recovery

A.8.1 Group Testing

Here we characterize the asymptotic behavior of the logarithm in (40). The main step is to upper bound the summation in the denominator, which is given by $\sum_{j=0}^{\lfloor\alpha k\rfloor}\binom{p-L}{j}\binom{L}{k-j}$. By the assumption $L = o(p)$, the value $j = \lfloor\alpha k\rfloor$ must yield the highest value of $\binom{p-L}{j}\binom{L}{k-j}$ when p is sufficiently large. Hence, upper bounding the summation by $\alpha k + 1$ times the maximum yields $\sum_{j=0}^{\lfloor\alpha k\rfloor}\binom{p-L}{j}\binom{L}{k-j} \le (\alpha k+1)\binom{p-L}{\lfloor\alpha k\rfloor}\binom{L}{k-\lfloor\alpha k\rfloor}$. Applying $\log\binom{a}{b} \le b\log\frac{ae}{b}$, we deduce that

$$\log\sum_{j=0}^{\lfloor\alpha k\rfloor}\binom{p-L}{j}\binom{L}{k-j} \le \log(\alpha k+1) + \lfloor\alpha k\rfloor\log\frac{pe}{\lfloor\alpha k\rfloor} + (k-\lfloor\alpha k\rfloor)\log\frac{Le}{k-\lfloor\alpha k\rfloor}. \tag{140}$$

Since $\log\frac{p}{k} \to \infty$ and $\alpha \in (0,1)$ does not depend on p, a simple asymptotic analysis yields $\log\sum_{j=0}^{\lfloor\alpha k\rfloor}\binom{p-L}{j}\binom{L}{k-j} \le \big(\alpha k\log\frac{p}{k} + (1-\alpha)k\log\frac{L}{k}\big)(1+o(1))$. The logarithm in (40) therefore simplifies to $\big(k\log\frac{p}{L}\big)(1+o(1))$, as desired.

A.8.2 Graphical Model Selection

Here we upper bound the quantity $N_{\max}(\alpha p)$ in (54). For all j, the first combinatorial term is upper bounded by 2^p, the second is maximized by $j = \alpha p$ (for sufficiently large p), and further upper bounding $\binom{p}{2} - p + 1 \le p^2$ yields $N_{\max}(\alpha p) \le (\alpha p+1)\cdot 2^p \cdot \binom{p^2}{\alpha p}$. Taking the logarithm and applying $\log\binom{a}{b} \le a\log\frac{ae}{b}$ along with asymptotic simplifications, we find that $N_{\max}(\alpha p) \le \big(\alpha p\log p\big)(1+o(1))$, as desired.

A.9 Proof of Theorem 10 (Reduction to Approximate Recovery)

We adopt the same general approach as Theorem 9, but instead of the error probability $\mathbb{P}_v[\hat{V} \neq v]$, we consider an approximate recovery version of the form $\mathbb{P}_v[d(v, \hat{V}) > t]$. We again start with (66), which we repeat here:

$$\sup_{\theta \in \Theta} \mathbb{E}_\theta \big[\ell(\theta, \hat{\theta})\big] = \Phi(\epsilon_0) \sup_{\theta \in \Theta} \mathbb{P}_\theta[\rho(\theta, \hat{\theta}) \geq \epsilon_0], \tag{141}$$

for any $\epsilon_0 > 0$. Consider the following minimum-distance rule for $\hat{V}$:

$$\hat{V} = \operatorname*{arg\,min}_{v=1,\ldots,M} \rho(\theta_v, \hat{\theta}), \tag{142}$$

and suppose that we have $\rho(\theta_v, \hat{\theta}) < \frac{\epsilon}{2}$ for the correct index v. Then for any $v' \in \mathcal{V}$ such that $d(v, v') > t$, we have

$$\rho(\theta_{v'}, \hat{\theta}) \geq \rho(\theta_v, \theta_{v'}) - \rho(\theta_v, \hat{\theta}) \tag{143}$$

$$> \epsilon - \frac{\epsilon}{2} = \frac{\epsilon}{2}, \tag{144}$$

where (143) follows from the triangle inequality, and (144) follows from (72) and the assumption $\rho(\theta_v, \hat{\theta}) < \frac{\epsilon}{2}$. As a result, when $\rho(\theta_v, \hat{\theta}) < \frac{\epsilon}{2}$, the minimum-distance rule (142) must output some $\hat{v}$ satisfying $d(v, \hat{v}) \leq t$, yielding

$$\mathbb{P}_v\left[\rho(\theta_v, \hat{\theta}) \geq \frac{\epsilon}{2}\right] \geq \mathbb{P}_v[d(v, \hat{V}) > t]. \tag{145}$$

With the above tools in place, we proceed as follows:

$$\sup_{\theta \in \Theta} \mathbb{P}_\theta\left[\rho(\theta, \hat{\theta}) \geq \frac{\epsilon}{2}\right] \geq \max_{v=1,\ldots,M} \mathbb{P}_v\left[\rho(\theta_v, \hat{\theta}) \geq \frac{\epsilon}{2}\right] \tag{146}$$

$$\geq \max_{v=1,\ldots,M} \mathbb{P}_v[d(v, \hat{V}) > t] \tag{147}$$

$$\geq \frac{1}{M} \sum_{v=1,\ldots,M} \mathbb{P}_v[d(v, \hat{V}) > t] \tag{148}$$

$$\geq 1 - \frac{I(V; \mathbf{Y}) + \log 2}{\log \frac{M}{N_{\max}(t)}}, \tag{149}$$

where (146) follows by maximizing over a smaller set, (147) follows from (145), (148) lower bounds the maximum by the average, and (149) follows from Fano's inequality for approximate recovery (*cf.*, Theorem 2) and the fact that $I(V; \hat{V}) \leq I(V; \mathbf{Y})$ by the data processing inequality (*cf.*, Lemma 1). The proof of (73) is concluded by substituting (149) into (141) with $\epsilon_0 = \frac{\epsilon}{2}$, and taking the infimum over all estimators $\hat{\theta}$.

A.10 Proof of Theorem 11 (Reduction for Noisy Optimization)

We follow a similar proof to that of Theorem 9, which gave an analogous result for estimation. First, by Markov's inequality, we have

$$\sup_{f \in \mathcal{F}} \mathbb{E}_f[\ell_f(\hat{X})] \geq \sup_{f \in \mathcal{F}} \epsilon \cdot \mathbb{P}_f[\ell_f(\hat{X}) \geq \epsilon]. \tag{150}$$

We proceed by analyzing the probability on the right-hand side.

Suppose that a random index V is drawn uniformly from $\{1, \ldots, M\}$, and the triplet $(\mathbf{X}, \mathbf{Y}, \hat{X})$ is generated

by running the optimization algorithm on f_V. Moreover, given $\hat{X} = \hat{x}$, let $\hat{V}$ index the function among $\{f_1, \ldots, f_M\}$ with the lowest corresponding value: $\hat{V} = \arg\min_{v=1,\ldots,M} f_v(\hat{x})$. By the assumption that any point x satisfies $\ell_{f_v}(x) \leq \epsilon$ at most one of the functions, we find that if $\ell_{f_v}(\hat{x}) \leq \epsilon$, then it must hold that $\hat{V} = v$. Hence, we have

$$\mathbb{P}_v\big[\ell_{f_v}(\hat{X}) > \epsilon\big] \geq \mathbb{P}_v[\hat{V} \neq v], \tag{151}$$

where $\mathbb{P}_v$ is a shorthand for $\mathbb{P}_{f_v}$.

With the above tools in place, we proceed as follows:

$$\sup_{f \in \mathcal{F}} \mathbb{P}_f\big[\ell_f(\hat{X}) \geq \epsilon\big] \geq \max_{v=1,\ldots,M} \mathbb{P}_v\big[\ell_{f_v}(\hat{X}) \geq \epsilon\big] \tag{152}$$

$$\geq \max_{v=1,\ldots,M} \mathbb{P}_v[\hat{V} \neq v] \tag{153}$$

$$\geq \frac{1}{M} \sum_{v=1,\ldots,M} \mathbb{P}_v[\hat{V} \neq v] \tag{154}$$

$$\geq 1 - \frac{I(V; \mathbf{X}, \mathbf{Y}) + \log 2}{\log M}, \tag{155}$$

where (152) follows by maximizing over a smaller set, (153) follows from (151), (154) lower bounds the maximum by the average, and (155) follows from Fano's inequality (*cf.*, (8) in Theorem 1) and the fact that $I(V; \hat{V}) \leq I(V; \mathbf{X}, \mathbf{Y})$ by the data processing inequality (*cf.*, third part of Lemma 3). The proof of (79) is concluded by substituting (155) into (150) and taking the infimum over all $\hat{X}$. For $M = 2$, we obtain (64) in the same way upon replacing (155) by the version of Fano's inequality for $M = 2$ given in Remark 1.

Acknowledgments

J. Scarlett was supported by an NUS startup grant. V. Cevher was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement 725594 – time-data).

References

[1] R. M. Fano, "Class notes for MIT course 6.574: Transmission of information," 1952.

[2] M. Malyutov, "The separating property of random matrices," *Math. Notes Acad. Sci. USSR*, vol. 23, no. 1, pp. 84–91, 1978.

[3] G. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1880–1901, March 2012.

[4] M. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, Dec. 2009.

[5] E. J. Candes and M. A. Davenport, "How well can we estimate a sparse vector?" *Appl. Comp. Harm. Analysis*, vol. 34, no. 2, pp. 317–323, 2013.

[6] H. Hassanieh, P. Indyk, D. Katabi, and E. Price, "Nearly optimal sparse Fourier transform," in *ACM Symp. Theory Comp. (STOC)*, 2012, pp. 563–578.

[7] V. Cevher, M. Kapralov, J. Scarlett, and A. Zandieh, "An adaptive sublinear-time block sparse Fourier transform," in *ACM Symp. Theory Comp. (STOC)*, 2017.

[8] A. A. Amini and M. J. Wainwright, "High-dimensional analysis of semidefinite relaxations for sparse principal components," *Annals Stats.*, vol. 37, no. 5B, pp. 2877–2921, 2009.

[9] V. Q. Vu and J. Lei, "Minimax rates of estimation for sparse PCA in high dimensions," in *Int. Conf. Art. Intel. Stats. (AISTATS)*, 2012, pp. 1278–1286.

[10] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *J. Mach. Learn. Res.*, vol. 13, pp. 1665–1697, May 2012.

[11] M. A. Davenport, Y. Plan, E. Van Den Berg, and M. Wootters, "1-bit matrix completion," *Information and Inference*, vol. 3, no. 3, pp. 189–223, 2014.

[12] I. A. Ibragimov and R. Z. Khasminskii, "Estimation of infinite-dimensional parameter in Gaussian white noise," *Doklady Akademii Nauk SSSR*, vol. 236, no. 5, pp. 1053–1055, 1977.

[13] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Annals. Stats.*, vol. 27, no. 5, pp. 1564–1599, 1999.

[14] L. Birgé, "Approximation dans les espaces métriques et théorie de l'estimation," *Prob. Theory and Related Fields*, vol. 65, no. 2, pp. 181–237, 1983.

[15] G. Raskutti, M. J. Wainwright, and B. Yu, "Minimax-optimal rates for sparse additive models over kernel classes via convex programming," *J. Mach. Learn. Res.*, vol. 13, no. Feb, pp. 389–427, 2012.

[16] Y. Yang, M. Pilanci, and M. J. Wainwright, "Randomized sketches for kernels: Fast and optimal non-parametric regression," *Annals Stats.*, vol. 45, no. 3, pp. 991–1023, 2017.

[17] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright, "Information-theoretic lower bounds for distributed statistical estimation with communication constraints," in *Conf. Neur. Inf. Proc. Sys. (NIPS)*, 2013, pp. 2328–2336.

[18] A. Xu and M. Raginsky, "Information-theoretic analysis of generalization capability of learning algorithms," in *Conf. Neur. Inf. Proc. Sys.*, 2017.

[19] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, "Local privacy and statistical minimax rates," in *IEEE Symp. Found. Comp. Sci. (FOCS)*, 2013.

[20] M. Raginsky and A. Rakhlin, "Information-based complexity, feedback and dynamics in convex programming," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 7036–7056, Oct. 2011.

[21] A. Agarwal, P. L. Bartlett, P. Ravikumar, and M. J. Wainwright, "Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3235 – 3249, May 2012.

[22] M. Raginsky and A. Rakhlin, "Lower bounds for passive and active learning," in *Conf. Neur. Inf. Proc. Sys. (NIPS)*, 2011.

[23] A. Agarwal, S. Agarwal, S. Assadi, and S. Khanna, "Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons," in *Conf. Learn. Theory (COLT)*, 2017.

[24] J. Scarlett, "Tight regret bounds for bayesian optimization in one dimension," in *Int. Conf. Mach. Learn. (ICML)*, 2018.

[25] Z. Bar-Yossef, T. S. Jayram, R. Kumar, and D. Sivakumar, "Information theory methods in communication complexity," in *IEEE Conf. Comp. Complex.*, 2002, pp. 93–102.

[26] N. Santhanam and M. Wainwright, "Information-theoretic limits of selecting binary graphical models in high dimensions," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4117–4134, July 2012.

[27] K. Shanmugam, R. Tandon, A. Dimakis, and P. Ravikumar, "On the information theoretic limits of learning Ising models," in *Adv. Neur. Inf. Proc. Sys. (NIPS)*, 2014.

[28] N. B. Shah and M. J. Wainwright, "Simple, robust and optimal ranking from pairwise comparisons," *J. Mach. Learn. Res.*, vol. 18, no. 199, pp. 1–38, 2018.

[29] A. Pananjady, C. Mao, V. Muthukumar, M. J. Wainwright, and T. A. Courtade, "Worst-case vs average-case design for estimation from fixed pairwise comparisons," 2017, http://arxiv.org/abs/1707.06217.

[30] Y. Yang, "Minimax nonparametric classification. i. rates of convergence," *IEEE Trans. Inf. Theory*, vol. 45, no. 7, pp. 2271–2284, 1999.

[31] M. Nokleby, M. Rodrigues, and R. Calderbank, "Discrimination on the Grassmann manifold: Fundamental limits of subspace classifiers," *IEEE Trans. Inf. Theory*, vol. 61, no. 4, pp. 2133–2147, April 2015.

[32] A. Mazumdar and B. Saha, "Query complexity of clustering with side information," in *Conf. Neur. Inf. Proc. Sys. (NIPS)*, 2017.

[33] E. Mossel, "Phase transitions in phylogeny," *Trans. AMS*, vol. 356, no. 6, pp. 2379–2404, 2004.

[34] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 2006.

[35] J. C. Duchi and M. J. Wainwright, "Distance-based and continuum Fano inequalities with applications to statistical estimation," 2013, http://arxiv.org/abs/1311.2669.

[36] I. Sason and S. Verdú, "f-divergence inequalities," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 5973–6006, Nov. 2016.

[37] R. Dorfman, "The detection of defective members of large populations," *Ann. Math. Stats.*, vol. 14, no. 4, pp. 436–440, 1943.

[38] J. Scarlett and V. Cevher, "Phase transitions in group testing," in *Proc. ACM-SIAM Symp. Disc. Alg. (SODA)*, 2016.

[39] ——, "How little does non-exact recovery help in group testing?" in *IEEE Int. Conf. Acoust. Sp. Sig. Proc. (ICASSP)*, 2017.

[40] L. Baldassini, O. Johnson, and M. Aldridge, "The capacity of adaptive group testing," in *IEEE Int. Symp. Inf. Theory*, July 2013, pp. 2676–2680.

[41] J. Scarlett and V. Cevher, "Converse bounds for noisy group testing with arbitrary measurement matrices," in *IEEE Int. Symp. Inf. Theory*, Barcelona, 2016.

[42] ——, "On the difficulty of selecting Ising models with approximate recovery," *IEEE Trans. Sig. Inf. Proc. over Networks*, vol. 2, no. 4, pp. 625–638, 2016.

[43] ——, "Lower bounds on active learning for graphical model selection," in *Int. Conf. Art. Intel. Stats. (AISTATS)*, 2017.

[44] V. Y. F. Tan, A. Anandkumar, and A. S. Willsky, "Learning high-dimensional Markov forest distributions: Analysis of error rates," *J. Mach. Learn. Res.*, vol. 12, no. May, pp. 1617–1653, 2011.

[45] A. Anandkumar, V. Y. F. Tan, F. Huang, and A. S. Willsky, "High-dimensional structure estimation in Ising models: Local separation criterion," *Ann. Stats.*, vol. 40, no. 3, pp. 1346–1375, 2012.

[46] G. Dasarathy, A. Singh, M.-F. Balcan, and J. H. Park, "Active learning algorithms for graphical model selection," in *Int. Conf. Art. Intel. Stats. (AISTATS)*, 2016.

[47] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*. Springer, 1997, pp. 423–435.

[48] J. Duchi, "Lecture notes for statistics 311/electrical engineering 377 (MIT)," http://stanford.edu/class/stats311/.

[49] Y. Wu, "Lecture notes for ECE598YW: Information-theoretic methods for high-dimensional statistics," http://www.stat.yale.edu/~yw562/ln.html.

[50] Y. Polyanskiy, V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[51] O. Johnson, "Strong converses for group testing from finite blocklength results," *IEEE Trans. Inf. Theory*, vol. 63, no. 9, pp. 5923–5933, Sept. 2017.

[52] R. Venkataramanan and O. Johnson, "A strong converse bound for multiple hypothesis testing, with applications to high-dimensional estimation," *Elec. J. Stats.*, vol. 12, no. 1, pp. 1126–1149, 2018.

[53] P.-L. Loh, "On lower bounds for statistical learning theory," *Entropy*, vol. 19, no. 11, p. 617, 2017.

[54] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. App. Math.*, vol. 6, no. 1, pp. 4 – 22, 1985.

[55] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, "Gambling in a rigged casino: The adversarial multi-armed bandit problem," in *IEEE Conf. Found. Comp. Sci. (FOCS)*, 1995.

[56] E. Arias-Castro, E. J. Candes, and M. A. Davenport, "On the fundamental limits of adaptive sensing," *IEEE Trans. Inf. Theory*, vol. 59, no. 1, pp. 472–481, Jan. 2013.

[57] T. S. Han and S. Verdú, "Generalizing the Fano inequality," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1247–1251, 1994.

[58] L. Birgé, "A new lower bound for multiple hypothesis testing," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1611–1615, 2005.

[59] A. A. Gushchin, "On Fano's lemma and similar inequalities for the minimax risk," *Prob. Theory and Math. Stats.*, vol. 67, pp. 26–37, 2004.

[60] A. Guntuboyina, "Lower bounds for the minimax risk using f-divergences, and applications," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2386–2399, 2011.

[61] Y. Polyanskiy and S. Verdú, "Arimoto channel coding converse and rényi divergence," in *Allerton Conf. Comm., Control, Comp.*, 2010.

[62] G. Braun and S. Pokutta, "An information diffusion Fano inequality," http://arxiv.org/abs/1504.05492.

[63] X. Chen, A. Guntuboyina, and Y. Zhang, "On Bayes risk lower bounds," *J. Mach. Learn. Res.*, vol. 17, no. 219, pp. 1–58, 2016. [Online]. Available: http://jmlr.org/papers/v17/16-185.html