Minimaxity in Structured Normal Means Inference

Akshay Krishnamurthy *1

1Computer Science Department, Carnegie Mellon University

June 29, 2015

Abstract

We provide a unified treatment of a broad class of noisy structure recovery problems, known as structured normal means problems. In these problems, the goal is to identify, from a finite collection of Gaussian distributions with different means, the distribution that produced the observed data. Recent work has studied several special cases including sparse vectors, biclusters, and graph-based structures. We establish nearly matching upper and lower bounds on the minimax probability of error for any structured normal means problem. We also derive an optimality certificate for the maximum likelihood estimator, which can be applied to many instantiations. We consider an experimental design setting, where we generalize our minimax bounds and derive an algorithm for computing a design strategy with a certain optimality property. We show that our results give tight minimax bounds for many structure recovery problems and consider some consequences for interactive sampling.

1 Introduction

The prevalence of high-dimensional signals in modern scientific investigation has inspired an influx of research on recovering structural information from noisy data. These problems arise across a variety of scientific and engineering disciplines; for example identifying cluster structure in communication or social networks, multiple hypothesis testing in genomics, or anomaly detection in vision and sensor networking. Broadly speaking, this line of work shows that high-dimensional statistical inference can be performed at low signal-to-noise ratios provided that the data exhibits low-dimensional structure. Specific structural assumptions include sparsity [13], low-rankedness [12], cluster structure [16], and many others [7].

The literature in this direction focuses on three inference goals: detection, localization or recovery, and estimation or denoising. Detection tasks involve deciding whether an observation contains some meaningful information or is simply ambient noise, while recovery and estimation tasks involve more precisely characterizing the information contained in a signal. Specifically, in recovery problems, the goal is to identify, from a finite collection of signals, which signal produced the observed data. The estimation or denoising problem involves leveraging structural information to produce high-quality estimates of the signal generating the data. These problems are closely related, but also exhibit important differences, and this paper focuses on the recovery problem.

*akshaykr@cs.cmu.edu
One frustration among researchers is that algorithmic and analytic techniques for these problems differ significantly for different structural assumptions. This issue was recently resolved in the context of the estimation problem, where the atomic norm\footnote{7} has provided a unifying algorithmic and analytical framework, but such a theory for detection and recovery problems remains elusive. In this paper, we provide a unification for the recovery problem, giving us better understanding of how signal structure affects statistical performance.

Modern measurement technology also often provides flexibility in designing strategies for data acquisition, and this adds an element of complexity to inference tasks. As a concrete example, crowdsourcing platforms allow for interactive data acquisition, which can be used to recover cluster structure with lower measurement overhead\cite{22,17}. Non-interactive experimental design-based (i.e. non-uniform) data acquisition is also enabled by modern sensing technology, leading to two important questions: (1) How do we design sensing strategies for structure recovery problems? (2) When should interactive acquisition be preferred to non-interactive acquisition? We provide an answer to the first of these questions, and progress toward an answer to the latter.

To concretely describe our main contributions, we now develop the decision-theoretic framework of this paper. We study the structured normal means problem defined by a finite collection of vectors $\mathcal{V} = \{v_j\}_{j=1}^M \subset \mathbb{R}^d$ that index a family of probability distributions $\mathbb{P}_j = \mathcal{N}(v_j,I_d)$. An estimator $T$ for the family $\mathcal{V}$ is a measurable function from $\mathbb{R}^d$ to $[M]$, and its maximum risk is:

$$R(T,\mathcal{V}) = \sup_{j \in [M]} R_j(T,\mathcal{V}), \quad R_j(T,\mathcal{V}) = \mathbb{P}_j[T(y) \neq j],$$

where we always use $y \sim \mathbb{P}_j$ to be the observation. We are interested in the minimax risk:

$$R(\mathcal{V}) = \inf_T R(T,\mathcal{V}) = \inf_T \sup_{j \in [M]} \mathbb{P}_j[T(y) \neq j]. \quad (1)$$

We call this the isotropic setting because each gaussian has spherical covariance. We are specifically interested in understanding how the complexity of the family $\mathcal{V}$ influences the minimax risk. This setting encompasses recent work on sparsity recovery\cite{13}, biclustering\cite{16,6}, and many graph-based problems\cite{23}. An example to keep in mind is the $k$-sets problem, where the collection $\mathcal{V}$ is formed by vectors $\mu 1_S$ for subsets $S \subset [d]$ of size $k$ and some signal strength parameter $\mu$.

We also study the experimental design setting, where the learning algorithm can specify a sensing strategy, defined by a vector $B \in \mathbb{R}^d_+$. Using this strategy, under $\mathbb{P}_j$, the observation is:

$$y(i) \sim v_j(i) + B(i)^{-1/2} \mathcal{N}(0,1) = \mathcal{N}(v_j(i), B(i)^{-1}), \forall i \in [d]. \quad (2)$$

If $B(i) = 0$, then we say that $y(i) = 0$ almost surely. We call this distribution $\mathbb{P}_{j,B}$, to denote the dependence both on the target signal $v_j$ and the sensing strategy $B$. The total measurement effort, or budget, used by the strategy is $\|B\|_1$, and we are typically interested in signal recovery under some budget constraint. Specifically, the minimax risk in this setting is:

$$R(\mathcal{V},\tau) = \inf_{T,B : \|B\|_1 \leq \tau} \sup_{j \in [M]} \mathbb{P}_{j,B}[T(y) \neq j]. \quad (3)$$

With this formalization, we can now state our main contributions:

1. We give nearly matching upper and lower bounds on the minimax risk for both isotropic and experimental design settings (Theorems\footnote{2} and\footnote{7}). This result matches many special cases that we are aware of\cite{23}, which we show through examples. Moreover, in examples with an asymptotic flavor (defined below), this shows that the maximum likelihood estimator (MLE) achieves the minimax rate.
2. In the isotropic case, we derive a condition on the family $\mathcal{V}$ under which the MLE exactly achieves the minimax risk, which certifies optimality of this estimator. In this case, we also give a heuristic algorithm that exploits connections to Bayesian inference and attempts to improve on the MLE. This algorithm gives some insights into how to appropriately regularize an inference problem.

3. We give sufficient conditions that certify an optimality property of an experimental design strategy and also give an algorithm for computing such a strategy prior to data acquisition. We give an example where a non-uniform strategy outperforms the isotropic one and two examples (one well-known and one new) where interactive strategies provably outperform all non-interactive ones. This latter result shows that interactive sampling can be significantly more powerful than non-interactive experimental design.

2 Related Work

The structured normal means problem has a rich history in statistics, although the majority of work focus on detection or estimation in nonparametric settings, for example when the signals belong to Besov or Sobolev spaces. More recently attention has turned to combinatorial structures and the finite dimensional case. This line is motivated by statistical applications involving complex data sources, such as tasks in graph-structured signal processing, and the broad goal is to understand how combinatorial structures affect both statistics and computation in these inference problems.

Focusing on detection problems, a number of papers study various combinatorial structures, including $k$-sets, cliques, paths, and clusters in graphs, and for many of these problems, near-optimal detection is possible. For example, Addario-Berry et al. show that to test between the null hypothesis that every component of the vector is $N(0, 1)$ and the alternative that $k$ components have mean $\mu$, the detection threshold is $\mu \approx \sqrt{\log(1 + \frac{4}{d_k^2})}$. This means that if $\mu$ grows faster than this threshold, one can achieve error probability tending to zero, and if $\mu$ grows slower than this threshold, all procedures have error probability tending to 1. This style of result is now available for several examples, although a unifying theory for detection problems is still undeveloped.

Turning to recovery or localization, again several specific examples have been analyzed. The most popular example is the biclustering problem, where $\mathcal{V}$ corresponds to $d_1 \times d_2$ matrices of the form $\mu 1_{C_1} 1_{C_r}^T$ with $C_1 \subset [d_1], C_r \subset [d_2]$. However, apart from this example and a few others, minimax bounds for the recovery problem are largely unknown. Moreover, we are unaware of a broadly applicable analysis, like the method we develop here.

A unified treatment is possible for estimation problems, where the atomic norm framework gives sharp phase transitions for the maximum likelihood estimator. The atomic norm is a generic approach for encoding structural assumptions by decomposing the signal into a sparse convex combination of a set of base atoms (e.g., one-sparse vectors). While this line primarily focuses on linear inverse problems, there are results for the estimation problem described above. While we are unaware of minimax bounds for either setting, it is well known that the the mean squared error of the MLE is related to the statistical dimension of the cone formed by the atoms. Unfortunately, atomic norm techniques rely on convex relaxation which enables estimation but not recovery, as the minimax probability of error for any dense family is one. Moreover, the non-convexity of our risk poses new challenges that do not arise with the strongly-convex mean squared error objective.

While much of the literature has focused on the isotropic case, there has been recent interest in experimental design or interactive methods, aiming to quantify the statistical improvements enabled by interac-
tivity. The first result in this line is a simple interactive procedure for the \(k\)-sets recovery problem due to Haupt, Castro and Nowak [13]. More recently, Tanczos and Castro [23] study more structured instantiations and show more significant statistical improvements via interactive methods. Their work makes important progress, but it does not address the general problem, as they hand-craft sampling algorithms for each example. A unifying, interactive algorithm was proposed in the bandit optimization setting [8], but, in our setting, it is not known to improve on non-interactive approaches. To our knowledge, a unifying interactive algorithm and a satisfactory characterization of the advantages offered by interactive sampling remain elusive open questions. This paper makes progress on the latter by developing lower bounds against all non-interactive approaches.

Lastly, there is a close connection between our setting and the channel coding problem in an Additive White Gaussian Noise (AWGN) Channel [10, 11]. In channel coding, we are tasked with designing a large code \(V\) such that if we send the codeword \(v_j\), an observer, upon observing \(y \sim N(v_j, I_d)\), can reliably predict the codeword sent. While the error metric is usually the same as in our setup, typical coding-theoretic results focus on codebook design, rather than error analysis for a particular codebook, which is our focus here. To our knowledge, the results here do not appear in the information theory literature.

### 3 Main Results

In this section we develop the main results of the paper. We start by bounding the minimax risk in the isotropic setting, then develop a certificate of optimality for the maximum likelihood estimator, and turn to the algorithmic question of computing minimax optimal estimators. Lastly, we turn to the experimental design setting. We provide proofs in Appendix A.

#### 3.1 Bounds on the Minimax Risk

In the isotropic case, recall that we are given a finite collection \(V\) of vectors \(\{v_j\}_{j=1}^M\) and an observation \(y \sim N(v_j, I_d)\) for some \(j \in [M]\). Given such an observation, a natural estimator is the maximum likelihood estimator (MLE), which outputs the index \(j\) for which the observation was most likely to have come from. This estimator is defined as:

\[
T_{\text{MLE}}(y) = \arg\max_{j \in [M]} P_j(y) = \arg\min_{j \in [M]} \|v_j - y\|^2_2. \tag{4}
\]

We will analyze this estimator, which partitions \(\mathbb{R}^d\) based on a Voronoi Tessellation of the set \(V\).

As stated, the running time of the estimator is \(O(Md)\), but it is worth pausing to remark briefly about computational considerations. In many examples of interest, the class \(V\) is combinatorial in nature, so \(M\) may be exponentially large, and efficient implementations of the MLE may not exist. However, as our setup does not preclude unstructured problems, the input to the estimator is the complete collection \(V\), so the running time of the MLE is linear in the input size. If the particular problem is such that \(V\) can be compactly represented (e.g. it has combinatorial structure), then the estimator may not be polynomial-time computable. This presents a real issue, as researchers have shown that a minimax-optimal polynomial time estimator is unlikely to exist for the biclustering problem [9, 13], which we study in Section 4. However, since the primary interest of this work is statistical in nature, we will ignore computational considerations for most of our discussion.

We now turn to a characterization of the minimax risk, which involves analysis of the MLE. The following function, which we call the **Exponentiated Distance Function**, plays a fundamental role.
Definition 1. For a family $\mathcal{V}$ and $\alpha > 0$, the Exponentiated Distance Function (EDF) is:

$$W(\mathcal{V}, \alpha) = \max_{j \in [M]} W_j(\mathcal{V}, \alpha) \quad \text{with} \quad W_j(\mathcal{V}, \alpha) = \sum_{k \neq j} \exp \left( \frac{-\|v_j - v_k\|^2}{\alpha} \right)$$ (5)

In the following theorem, we show that the EDF governs the performance of $T_{\text{MLE}}$. More importantly, this function also leads to a lower bound on the minimax risk, and the combination of these two statements shows that the MLE is nearly optimal for any structured normal means problem.

Theorem 2. Fix $\delta \in (0, 1)$. If $W(\mathcal{V}, 8) \leq \delta$, then $\mathcal{R}(\mathcal{V}) \leq \mathcal{R}(\mathcal{V}, T_{\text{MLE}}) \leq \delta$. On the other hand, if $W(\mathcal{V}, 2(1 - \delta)) \geq 2^{\frac{1}{1-\delta}} - 1$, then $\mathcal{R}(\mathcal{V}) \geq \delta$.

In particular, by setting $\delta = 1/2$ above, the second statement in the theorem may be replaced by: If $W(\mathcal{V}, 1) \geq 3$, then $\mathcal{R}(\mathcal{V}) \geq 1/2$. This setting often aids interpretability of the lower bound.

Notice that the value of $\alpha$ disagrees between the lower and upper bounds, and this leads to a gap between the necessary and sufficient conditions. This is not purely an artifact of our analysis, as there are many examples where the MLE does not exactly achieve the minimax risk. However, most structured normal means problems also have an asymptotic flavor, specified by a sequence of problems $\mathcal{V}_1, \mathcal{V}_2, \ldots$, and a signal-strength parameter $\mu$, with observation $y \sim N(\mu v_j, I_d)$ for some signal $v_j$ in the current family. In this asymptotic framework, we are interested in how $\mu$ scales with the sequence to drive the minimax risk to one or zero. Almost all existing examples in the literature are of this form [23], and in all such problems, Theorem 2 shows that the MLE achieves the minimax rate. To our knowledge, such a comprehensive characterization of recovery problems is entirely new.

Application of Theorem 2 to instantiations of the structured normal means problem requires bounding the EDF, which is significantly simpler than the typical derivation of this style of result. In particular, proving a lower bound no longer requires construction of a specialized subfamily of $\mathcal{V}$ as was the de facto standard in this line of work [16, 23]. In Section 3, we show how simple calculations can recover existing results.

Turning to the proof, the EDF arises naturally as an upper bound on the failure probability of the MLE after applying a union bound and a Gaussian tail bound. Indeed the fact that the EDF upper bounds the minimax risk is not particularly surprising. It is however more surprising that it also provides a lower bound on the minimax risk. We obtain this bound via application of Fano’s Inequality, but we use a version that allows a non-uniform prior and explicitly construct this prior using the EDF. This leads to our more general lower bound.

### 3.2 Minimax-Optimal Recovery

Theorem 2 shows that that maximum likelihood estimator achieves near-optimal performance for all structured normal means recovery problems. By near-optimal, we mean that in problems with some asymptotic flavor, where the family of vectors grows but also becomes more separated, the maximum likelihood estimator achieves the minimax rate. However, in many cases the MLE is not the optimal estimator, i.e. it does not achieve the exact minimax risk. In this section, we use deeper connections between the minimax risk and the Bayes risk to address this gap. Specifically, we give a sufficient condition for the minimax optimality of the MLE, and we will also design an algorithm that in other cases produces an estimator with better minimax risk.

Our approach is based on a well-known connection between the minimax risk and the Bayes risk. For a structured normal means problem defined by a family $\mathcal{V}$, the Bayes risk for an estimator $\hat{T}$ under prior
\( \pi \in \Delta_{M-1} \) is given by:

\[
\mathcal{B}_\pi(T) = \sum_{j=1}^{M} \pi_j \mathbb{P}_j [T(y) \neq j].
\]

We say that an estimator \( T \) is the Bayes estimator for prior \( \pi \) if it achieves the minimum Bayes risk. A simple calculation reveals the structure of the Bayes estimator for any prior \( \pi \) and this structural characterization is essential to our development.

**Proposition 3.** For any prior \( \pi \), the Bayes estimator \( T_\pi \) has polyhedral acceptance regions, that is the estimator is of the form:

\[
T(Y) = j \text{ if } y \in A_j,
\]

with \( A_j = \{ x : \Gamma_j x \geq b_j \} \) and \( \Gamma_j \in \mathbb{R}^{M \times d} \) has \( v_j - v_k \) in the \( k \)th row and \( b_j \) has \( \frac{1}{2} (\|v_j\|_2^2 - \|v_k\|_2^2) + \log \frac{\pi_j}{\pi_k} \) in the \( k \)th entry. These polyhedral sets \( A_j \) partition the space \( \mathbb{R}^d \).

We also exploit the relationship between the minimax risk and the Bayes risk. This is a well known result, although for completeness we provide a proof in Appendix A.4. The prior \( \pi \) below is known as the least-favorable prior.

**Proposition 4.** Suppose that \( T \) is a Bayes estimator for some prior \( \pi \). If the risk \( \mathcal{R}_j(T) = \mathcal{R}_{j'}(T) \) for all \( j \neq j' \in [M] \), then \( T \) is a minimax optimal estimator.

Our main theoretical result leverages this proposition along with the structural characterization of Bayes estimators to certify minimax optimality of the MLE. The sufficient condition for optimality depends on a particular structure of the family \( \mathcal{V} \):

**Definition 5.** A family \( \mathcal{V} \) is unitarily invariant if there exists a set of orthogonal matrices \( \{ R_i \}_{i=1}^{N} \) such that for each vector \( v \in \mathcal{V} \), the set \( \{ R_i v \}_{i=1}^{N} \) is exactly \( \mathcal{V} \).

In other words, the instance \( \mathcal{V} \) can be generated by applying the orthogonal transforms to any fixed vector in the collection. Unitarily invariant problems exhibit high degrees of symmetry and via Proposition 4, can be shown to be a sufficient condition for the optimality of the MLE.

**Theorem 6.** If \( \mathcal{V} \) is unitarily invariant, then the MLE is minimax optimal.

Some remarks about the theorem are in order:

1. This theorem reduces the question of optimality to a purely geometric characterization of the family \( \mathcal{V} \) and, as we will see, many well studied problems are unitarily invariant. One common family of orthogonal matrices is the set of all permutation matrices on \( \mathbb{R}^d \).

2. This result does not characterize the risk of the MLE; it only shows that no other estimator has better risk. Specifically, it does not provide an analytic bound that is sharper than Theorem 2. From a practitioner’s perspective, an optimality certificate for an estimator is more important than a bound on the risk as it help govern practical decisions, although risk bounds enable theoretical comparison.

3. Lastly, the result is not asymptotic in nature but rather shows that the MLE achieves the exact minimax risk for a fixed family \( \mathcal{V} \). We are not aware of any other results in the literature that certify optimality of the MLE under our measure of risk.
The proof of this theorem is based on the observation that the point-wise risk \( R_j(\mathcal{V}) \) is exactly \( 1 - P_j[A_j] \) where \( P_j \) is the gaussian measure centered at \( v_j \) and \( A_j \) is a particular polytope based on the Voronoi Tessellation of the point set \( \mathcal{V} \). We use this characterization and the unitary invariance of the family to show that the risk landscape for the MLE is constant across the hypotheses \( v_j \). Finally, we employ a dual characterization of the minimax risk, to show that if the risk landscape is constant, then the MLE must be optimal.

In problems where Theorem 4 can be applied, we now have a complete story for the isotropic case. We know that the MLE exactly achieves the minimax risk and Theorem 2 also gives satisfactory upper and lower bounds. However, many problems of interest do not have unitarily invariant structure, and, in many of these problems, the MLE is suboptimal.

### 3.3 The Experimental Design Setting

Recall the experimental design setting, where the statistician specifies a strategy \( B \in \mathbb{R}^d_+ \) and receives observation \( y \sim P_{j,B} \) given by Equation 2. Our main insight is that the choice of \( B \) only changes the metric structure of \( \mathbb{R}^d \), and this change can be incorporated into the proof of Theorem 2. Specifically, the likelihood for hypothesis \( j \), under sampling strategy \( B \) is:

\[
      P_j(y|B) = \prod_{i=1}^{d} \left( \frac{B(i)}{2\pi} \right)^{\frac{1}{2}} \exp\left( -\frac{1}{2} \sum_{i=1}^{d} B(i) (v_j(i) - y(i))^2 \right)
\]

and the maximum likelihood estimator is:

\[
      T_{MLE}(y, B) = \arg\min_{j \in [M]} ||v_j - y||_B^2
\]

where \( ||v||_B^2 = \sum_{i=1}^{d} v(i)^2 B(i) \) is the Mahalanobis norm induced by the diagonal matrix \( \operatorname{diag}(B) \).

Theorem 2 can be ported directly to this setting, leading to the following:

**Theorem 7.** Fix \( \delta \in (0, 1) \) and any sampling strategy \( B \) with \( ||B||_1 \leq \tau \). Define the Sampling Exponentiated Distance Function (SEDF):

\[
      W(\mathcal{V}, \alpha, B) = \max_{j \in [M]} \sum_{k \neq j} \exp\left( -\frac{||v_j - v_k||^2_2}{\alpha B} \right)
\]

If \( W(\mathcal{V}, 8, B) \leq \delta \) then \( R(\mathcal{V}, \tau) \leq R(\mathcal{V}, T_{MLE}(y, B)) \) \( \leq \delta \). Conversely, if \( W(\mathcal{V}, 2(1 - \delta), B) \geq 2^{\frac{1}{1-\delta}} - 1 \), then \( \inf_{\mathcal{V}} \sup_{j \in [M]} P_{j,B}[T(y) \neq j] \geq \delta \).

The structure of the theorem is almost identical to that of Theorem 2, but it is worth making some important observations. First, the theorem holds for any non-interactive sampling strategy \( B \in \mathbb{R}^d_+ \), so the upper bound is strictly more general than Theorem 2. Secondly, any non-interactive strategy can be used to derive an upper bound on the minimax risk, but the same is not true for the lower bound. Instead the lower bound provided by the theorem is dependent on the strategy, so one must still minimize over sampling strategies to lower bound \( R(\mathcal{V}, \tau) \). Note that this theorem also applies to the non-isotropic or heteroscedastic case with known, shared covariance.

Fortunately, the SEDF is convex in \( B \) so it can be numerically minimized over the polyhedron \( \{ z : 0 \leq z_i \leq 1, \sum_{i=1}^{d} z_i \leq B \} \). Specifically, for any \( \alpha \), we solve the convex program:

\[
      \min_{B \in \mathbb{R}^d_+, ||B||_1 \leq \tau} \max_{j \in [M]} \sum_{k \neq j} \exp\left( -\frac{||v_j - v_k||^2_2}{\alpha B} \right),
\]
To obtain the sampling strategy \( \hat{B} \) that minimizes the SEDF. For example, solving Program 7 with \( \alpha = 1 \) results in a strategy \( \hat{B} \), and if \( W(\mathcal{V}, 1, \hat{B}) \geq 3 \), then we know that the minimax risk \( R(\mathcal{V}, \tau) \) over all strategies is at least \( 1/2 \). On the other hand, solving with \( \alpha = 8 \) to obtain a (different) sampling strategy \( \hat{B} \) and then using \( \hat{B} \) with the MLE would give the tightest upper bound on the risk attainable by our proof technique. In Section 4, we demonstrate an example where this optimization leads to a non-uniform sampling strategy that outperforms uniform sampling.

In the general setting, it is challenging to analytically certify that an allocation strategy \( \hat{B} \) minimizes the SEDF, but in some cases it is possible. Specializing the first-order optimality conditions for Program 7 to our setting gives the following:

**Proposition 8.** Let \( \hat{B} \) be an sampling strategy with \( \|B\|_1 = \tau \). Let \( S(\hat{B}) \subset \mathcal{V} \) be the set of hypotheses achieving the maximum in \( W(\mathcal{V}, \alpha, \hat{B}) \) and let \( \pi \) be a distribution on \( S(\hat{B}) \). If, for all \( i, i' \in [d] \),

\[
\mathbb{E}_{j \sim \pi} \sum_{k \neq j} (v_k(i) - v_j(i))^2 \exp\left(-\|v_k - v_j\|_B^2\right) = \mathbb{E}_{j \sim \pi} \sum_{k \neq j} (v_k(i') - v_j(i'))^2 \exp\left(-\|v_k - v_j\|_B^2\right),
\]

then \( \hat{B} \) is a minimizer of \( W(\mathcal{V}, \alpha, B) \) subject to \( \|B\|_1 \leq \tau \).

While application of this result could involve a number of non-trivial calculations, there are many cases where it does lead to analytic lower bounds for particular problems. Specifically, the result is especially useful when \( \hat{B} \) is uniform across the coordinates, and \( S(\hat{B}) = [M] \), so that all of the hypotheses achieve the maximum. In this case, it often suffices to choose \( \pi \) to be uniform over the hypotheses and exploit the high degree of symmetry to demonstrate the condition holds. As we will see in Section 4, many examples studied in the literature exhibit the requisite symmetry for this proposition to be applied in a straightforward manner.

We remark that Tanczos and Castro [23] establish a similar sufficient condition for the uniform sampling strategy to be optimal. Their result however is slightly less general in that it only certifies optimality for the uniform sampling strategy, whereas ours, in principle, can be applied more universally. In addition, their result applies only to problems where the hypotheses are of the form \( \mu 1_S \) for a collection of subsets while ours is more general, and this generality is important for some examples (e.g., the hierarchical clustering example in Section 3). The other main difference is that their approach is not based on the SEDF, so their result is not directly applicable here.

### 4 Examples

To demonstrate the scope of our results, we present four instantiations of structured normal means problems, and derive results easily attainable from our general approach. These examples have the asymptotic flavor described before, where we are interested in how a signal strength parameter \( \mu \) scales with a sequence of problem instances. To simplify presentation, we state the results in terms of the minimax rate \( \psi \) and use the notation \( \mu \asymp \psi \) where \( \psi \) is a function that depends on the parameters of the sequence (e.g., the dimension). This notation means that if \( \mu = \omega(1) \psi \), then the minimax risk can be driven to zero and conversely, if \( \mu = o(1) \psi \), then the minimax risk approaches one asymptotically.

The first example, the \( k \)-sets problem, is well studied, and as a warmup, we show how our technique recovers existing results. The second example is the biclustering problem; this problem is interesting because there is polynomial separation between non-interactive and interactive procedures, and our technique can be used to establish lower bounds on all non-interactive approaches. The third example is a graph-structured
signal processing problem, and this example is interesting because our technique generalizes existing results, but also because uniform sampling may not be optimal. In the last example, we use Theorem 2 to demonstrate the achievability of the channel capacity in additive white gaussian noise (AWGN) channel, showing how an easy calculation can reproduce the proof of Shannon [20]. The requisite calculations for these examples are deferred to Appendix A.

4.1 k-sets

In the k-sets problem, we have \( M = \binom{d}{k} \) and each vector \( v_j = 1_{S_j} \) where \( S_j \subset [d] \) and \( |S_j| = k \). The observation is \( y \sim N(\mu v_j, I_d) \) for some hypothesis \( j \).

**Corollary 9.** The minimax rate for the k-sets problem is \( \mu \asymp \sqrt{\frac{d \log(k(d-k))}{k}} \) and with budget constraint \( \tau \), it is \( \mu \asymp \sqrt{\frac{d^2 \log(k(d-k))}{k}} \). In the isotropic case, the MLE is minimax optimal.

This corollary follows simply by bounding the EDF for the k-sets problem using binomial approximations. Using Proposition 5, it is easy to verify that uniform sampling is optimal here, which immediately gives the second claim. Finally using the set of all permutation matrices and exploiting symmetry, we can easily verify that this class is unitarily invariant. These bound agrees with established results in the literature [23].

4.2 Biclusters

In the biclustering problem, we instead work over \( \mathbb{R}^{d \times d} \) and let \( M = \binom{d}{k}^2 \). We parametrize the class \( \mathcal{V} \) with two indices so that \( v_{ij} = 1_{S_i}1_{S_j}^T \) is a \( d \times d \) matrix with \( k^2 \) non-zeros with \( |S_i| = |S_j| = k \). The observation is \( y \sim N(\mu \text{vec}(v_{ij}), I_{d^2}) \) for a hypothesis \( (i, j) \).

**Corollary 10.** The minimax rate for the biclustering problem is \( \mu \asymp \sqrt{\frac{\log(k(d-k))}{k}} \) and with budget constraint \( \tau \), it is \( \mu \asymp \sqrt{\frac{d^2 \log(k(d-k))}{k}} \). In the isotropic case, the MLE is minimax optimal.

Our bounds agree with existing analyses of this class [16] [6] [23]. Obtaining this result involves simply bounding the EDF using binomial approximations as in the k-sets example, and straightforward applications of Theorem 5 and Proposition 8 with the uniform distribution.

The biclustering problem is interesting because a simple interactive algorithm has significantly better statistical performance. The algorithm first samples coordinates of the matrix randomly, with enough energy so as to reliably test if a coordinate is active or not, until it finds an active coordinate. It then senses on the row and column of that coordinate and identifies the rows and columns that are active in the bicluster. Tanczos and Castro [23] show that this algorithm succeeds if \( \mu = \omega(\sqrt{\frac{d^2}{\pi^2 k^2} + \frac{d}{k}}) \log d \), which is a factor of \( \sqrt{k} \) smaller than the lower bound established here, demonstrating concrete statistical gains from interactivity. Note that this separation is known [23]. We provide a crude by sufficient analysis of this interactive algorithm in Appendix A.

4.3 Stars

Let \( G = (V, E) \) be a graph and let the edges be numbered \( 1, \ldots, d \). The class \( \mathcal{V} \) is the set of all stars in the graph, that is the vector \( v_j \in \{0, 1\}^d \) is the indicator vector of all edges emanating from the \( j \)th node in the graph. Again the observation is \( y \sim N(\mu v_j, I_d) \) for some \( j \in |V| \).
Figure 1: Left: A realization of the stars problem for a graph with 13 vertices and 34 edges with sampling budget $\tau = 34$. Edge color reflects allocation of sensing energy and vertex color reflects success probability for MLE under that hypothesis (warmer colors are higher for both). Isotropic (left) has minimum success probability of 0.44 and experimental design (center) has minimum success probability 0.56. Right: Maximum risk for isotropic and experimental design sampling as a function of $\mu$ for stars problem on a 50 and 100-vertex graph.

**Corollary 11.** In the stars problem if the ratio between the maximum and minimum degree is bounded by a constant, i.e. $\frac{\deg_{\text{max}}}{\deg_{\text{min}}} \leq c$, then the minimax rate is $\mu \approx \sqrt{\frac{\log(|V| - \deg_{\text{min}})}{\deg_{\text{min}}}}$.

Again this agrees with a recent result of Tanczos and Castro [23], who consider $s$-stars of the complete graph, formed by choosing a vertex, and then activating $s$ of the edges emanating out of that vertex. The two bounds agree in the special case of the complete graph with $s = |V| - 1$, but otherwise are incomparable, as they consider different problem structures. Note that the degree requirement here is not fundamental in Theorem 2 but rather a shortcoming of our calculations.

We highlight this example because the uniform allocation strategy does not necessarily minimize $W(V, \alpha, B)$. In Figure 1, we construct a graph according to the Barabási-Albert model [2] and consider the class of stars on this graph. The simulation results show that optimizing the SEDF to find a sampling strategy is never worse than uniform sampling, and for low signal strengths it can lead to significantly lower maximum risk. We believe the increases in the right-most plot are caused by numerical instability arising from the non-smooth optimization problem [7]. Note that the risk for both uniform and non-uniform sampling approaches zero as $\mu \to \infty$, so for large $\mu$, there is little advantage to optimizing the sampling scheme.

### 4.4 Random Codes

Consider a collection $\mathcal{V}$ of $M$ vectors with coordinates that are i.i.d. $\mathcal{N}(0, P)$. In expectation over the draw of the $M$ vectors, the Bayes risk of the maximum likelihood estimator under the uniform prior is:

$$\mathbb{E}_{\mathcal{V}}\mathbb{E}_{j \sim \text{Unif}[M]} P_j[\text{error}] \leq (M - 1)(1 + P/2)^{-d/2}$$

In other words if $M = o((1 + P/2)^{d/2})$, then the maximum likelihood decoder can drive the probability of error to zero as $d \to \infty$.

In information theoretic terms, this quick calculation roughly says that there exists a rate $R = \log(M)/d = \frac{1}{2} \log(1 + P/2) - \omega(1/d)$ code with power constraint $P$, that can be reliably transmitted over an additive white noise gaussian (AWGN) channel with noise variance 1. This nearly matches Shannon’s Channel
Coding theorem [10] which says that the rate cannot exceed the channel capacity, which in our case is \( \frac{1}{2} \log(1 + P) \).

There are two small weaknesses of this calculation in comparison with the classical achievability of the channel capacity. The first is that our bound involves the term \( \log(1 + P/2) \) instead of \( \log(1 + P) \) in the definition of channel capacity. We suspect this is due to weakness in our bounding technique in Theorem 2 which in part allows for significantly more generality than this special case. The second is that the codewords we use are drawn from \( N(0, P) \) so they will exceed the power constraint \( \|v\|^2 \leq P \) with constant probability. This shortcoming can be remedied by instead using \( N(0, P - c/d) \) and applying well known \( \chi^2 \) deviation bounds.

5 Discussion

In this paper, we studied the structured normal means problem and gave a unified characterization of the minimax risk both for isotropic and experimental design settings. Our work gives insights into how to choose estimators (e.g., the optimality certificate for the MLE) and how to design sampling strategies for structure recovery problems. Our lower bounds are critical in demonstrating separation between non-interactive and interactive sampling, which is an important research direction.

There are a number of exciting directions for future work, including extensions to other structure discovery problems such as detection, and to other observation models, such as compressive observations. We are most interested in developing a unifying theory for interactive sampling, analogous to the theory developed here. The challenges with developing such an understanding are both algorithmic and information theoretic, and we are excited to tackle these challenges.

Acknowledgements

AK would like to thank Aarti Singh, Sivaraman Balakrishnan, Martin Azizyan, and James Sharpnack for enlightening discussions during the development of this work.

References

[1] Louigi Addario-Berry, Nicolas Broutin, Luc Devroye, and Gábor Lugosi. On combinatorial testing problems. *The Annals of Statistics*, 2010.

[2] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 2002.

[3] Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference*, 2014.

[4] Ery Arias-Castro and Emmanuel J. Candès. Searching for a trail of evidence in a maze. *The Annals of Statistics*, 2008.

[5] Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. Atomic norm denoising with applications to line spectral estimation. *IEEE Transactions on Signal Processing*, 2013.
[6] Cristina Butucea and Yuri I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. Bernoulli, 2013.

[7] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. Foundations of Computational Mathematics, 2012.

[8] Shouyuan Chen, Tian Lin, Irwin King, Michael R. Lyu, and Wei Chen. Combinatorial Pure Exploration of Multi-Armed Bandits. In Advances in Neural Information Processing Systems, 2014.

[9] Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. arXiv:1402.1267, 2014.

[10] Thomas M. Cover and Joy A. Thomas. Elements of information theory. John Wiley & Sons, 2012.

[11] Imre Csiszar and János Körner. Information theory: coding theorems for discrete memoryless systems. Cambridge University Press, 2011.

[12] David Donoho and Matan Gavish. Minimax risk of matrix denoising by singular value thresholding. The Annals of Statistics, 2014.

[13] Jarvis Haupt, Rui Castro, and Robert Nowak. Distilled sensing: Adaptive sampling for sparse detection and estimation. IEEE Transactions on Information Theory, 2011.

[14] Yuri I. Ingster. Minimax testing of nonparametric hypotheses on a distribution density in the L_p metrics. Theory of Probability & Its Applications, 1987.

[15] Yuri I. Ingster and Irina A. Suslina. Nonparametric goodness-of-fit testing under Gaussian models. Springer-Verlag, 2003.

[16] Mladen Kolar, Sivaraman Balakrishnan, Alessandro Rinaldo, and Aarti Singh. Minimax Localization of Structural Information in Large Noisy Matrices. Advances in Neural Information Processing Systems, 2011.

[17] Akshay Krishnamurthy, Sivaraman Balakrishnan, Min Xu, and Aarti Singh. Efficient active algorithms for hierarchical clustering. In International Conference on Machine Learning, 2012.

[18] Zongming Ma and Yihong Wu. Computational barriers in minimax submatrix detection. arXiv:1309.5914, 2013.

[19] Samet Oymak and Babak Hassibi. Sharp mse bounds for proximal denoising. arXiv:1305.2714, 2013.

[20] Claude Elwood Shannon. Communication in the presence of noise. Proceedings of the IRE, 1949.

[21] James Sharpnack, Akshay Krishnamurthy, and Aarti Singh. Near-optimal Anomaly Detection in Graphs using Lovasz Extended Scan Statistic. Advances in Neural Information Processing Systems, 2013.

[22] Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. Adaptively learning the crowd kernel. arXiv:1105.1033, 2011.

[23] Ervin Tánczos and Rui Castro. Adaptive sensing for estimation of structured sparse signals. arXiv:1311.7118, 2013.
A Proofs

A.1 Proof of Theorem 2

Analysis of MLE: We first analyze the maximum likelihood estimator:

\[ T_{MLE}(y) = \arg\min_{j \in [M]} \|v_j - y\|^2_2 \]

This estimator succeeds as long as \( \|v_k - y\|^2_2 > \|v_{j^*} - y\|^2_2 \) for each \( k \neq j^* \), when \( y \sim P_{j^*} \). This condition is equivalent to:

\[ \|v_k - y\|^2_2 > \|v_{j^*} - y\|^2_2 \Leftrightarrow \langle \epsilon, v_k - v_{j^*} \rangle < \frac{1}{2}\|v_{j^*} - v_k\|^2_2 \]

where \( \epsilon \sim \mathcal{N}(0, 1) \). This follows from writing \( y = v_{j^*} + \epsilon \) and then expanding the squares. So we must simultaneously control all of these events, for fixed \( j^* \):

\[
\mathbb{P}_{\epsilon \sim \mathcal{N}(0,I_d)} \left[ \forall k \neq j^*, \langle \epsilon, v_k - v_{j^*} \rangle < \|v_{j^*} - v_k\|^2_2 / 2 \right] = 1 - \mathbb{P}_{\epsilon \sim \mathcal{N}(0,I_d)} \left[ \exists k \neq j^*, \langle \epsilon, v_k - v_{j^*} \rangle \geq \|v_{j^*} - v_k\|^2_2 / 2 \right] \\
\geq 1 - \sum_{k \neq j^*} \mathbb{P}_{\epsilon \sim \mathcal{N}(0,I_d)} \left[ \langle \epsilon, v_k - v_{j^*} \rangle \geq \|v_{j^*} - v_k\|^2_2 / 2 \right]
\]

By a gaussian tail bound, this probability is:

\[
\mathbb{P}_{\epsilon \sim \mathcal{N}(0,I_d)} \left[ \langle \epsilon, v_k - v_{j^*} \rangle \geq \|v_{j^*} - v_k\|^2_2 / 2 \right] \leq \exp \left\{ -\frac{1}{8} \|v_{j^*} - v_k\|^2_2 \right\}
\]

So that the total failure probability is upper bounded by:

\[
\mathbb{P}_{j^*} \left[ \hat{j} = j^* \right] \leq \sum_{k \neq j^*} \exp \left\{ -\frac{1}{8} \|v_{j^*} - v_k\|^2_2 \right\} = W_{j^*}(\mathcal{V}, 8)
\]

So if \( j \) is the truth, then the probability of error is smaller than \( \delta \) when \( W_{j}(\mathcal{V}, 8) \leq \delta \). For the maximal (over hypothesis choice \( j \)) probability of error to be smaller than \( \delta \), it suffices to have \( W(\mathcal{V}, 8) \leq \delta \).

Fundamental Limit: We now turn to the fundamental limit. We start with a version of Fano’s inequality with non-uniform prior.

Lemma 12 (Non-uniform Fano Inequality). Let \( \Theta = \{ \theta \} \) be a parameter space that indexes a family of probability distributions \( P_\theta \) over a space \( \mathcal{X} \). Fix a prior distribution \( \pi \) supported on \( \Theta \), and consider \( \theta \sim \pi \) and \( X \sim P_\theta \). Let \( f : \mathcal{X} \to \Theta \) be any possibly randomized mapping, and let \( p_e = \mathbb{P}_{\theta \sim \pi; X \sim P_\theta} [f(X) \neq \theta] \) denote the probability of error. Then:

\[
p_e \geq 1 - \sum_{\theta} \pi(\theta)KL(P_\theta || P_\pi) + \log 2
\]

where \( P_\pi(\cdot) = \mathbb{E}_{\theta \sim \pi} P_\theta(\cdot) \) is the mixture distribution. In particular, we have:

\[
\inf_{f} \sup_{\theta} \mathbb{P}_{X \sim P_\theta} [f(X) \neq \theta] \geq \inf_{f} \mathbb{E}_{\theta \sim \pi} \mathbb{P}_{X \sim P_\theta} [f(X) \neq \theta] \geq 1 - \sum_{\theta} \pi(\theta)KL(P_\theta || P_\pi) + \log 2
\]

\[ \frac{H(\pi)}{\mathbb{H}(\pi)} \]
Proof. Consider the Markov Chain \( \theta \rightarrow X \rightarrow \hat{\theta} \triangleq f(X) \) where \( \theta \sim \pi \) and \( X|\theta \sim P_\theta \). Let \( E = 1[\hat{\theta} \neq \theta] \).

\[
H(E|X) + H(\theta|E, X) = H(E, \theta|X) = H(\theta|X) + H(E|\theta, X) \geq H(\theta|X)
\]

Now, \( H(E|\theta, X) \geq 0 \) and since conditioning only reduces entropy, we have the inequality

\[
H(\theta|X) \leq H(p_e) + H(\theta|E, X) = H(p_e) + H(\theta|E = 0, X)P[E = 0] + H(\theta|E = 1, X)P[E = 1]
\]

\[
= H(p_e) + p_e H(\theta)
\]

which proves the usual version of Fano’s inequality. We want to write \( H(\theta|X) \) in terms of the KL divergence, using the mixture distribution \( P_{\pi} \).

\[
H(\theta|X) = H(\theta, X) - H(X) = \int \sum_{\theta} \pi(\theta) P_\theta(x) \log \left( \frac{\sum_{\theta} \pi(\theta) P_{\theta}(x)}{\pi(\theta) P_{\theta}(x)} \right) dx
\]

\[
= \sum_{\theta} \pi(\theta) \int P_\theta(x) \log \left( \frac{P_{\pi}(x)}{P_\theta(x)} \right) dx - \sum_{\theta} \pi(\theta) \log \pi(\theta)
\]

\[
= - \sum_{\theta} \pi(\theta) KL(P_\theta || P_{\pi}) + H(\pi)
\]

Combining these gives the bound:

\[
H(p_e) + p_e H(\pi) \geq H(\pi) - \sum_{\theta} \pi(\theta) KL(P_\theta || P_{\pi}),
\]

By upper bounding \( H(p_e) \leq \log 2 \) and rearranging we prove the claim.

For a distribution \( \pi \in \Delta_{M-1} \) over the hypothesis, let \( P_\pi(\cdot) = \sum_k \pi_k P_k(\cdot) \) be the mixture distribution. Then Fano’s inequality (Lemma 12) states that the minimax probability of error is lower bounded by:

\[
R(\mathcal{V}) = \inf_T \sup_j P_j[T(y) \neq j] \geq \inf_T \mathbb{E}_{y \sim \pi} \mathbb{E}_{y \sim \pi} 1[T(y) \neq j]
\]

\[
\geq 1 - \mathbb{E}_{k \sim \pi} KL(P_k || P_{\pi}) + \log 2
\]

\[
\frac{H(\pi)}{H(\pi)}.
\]

Fix \( \delta \in (0, 1) \) and let \( j^* = \arg\max_{j \in [M]} W_j(2(1 - \delta)) \). We will use a prior based on this quantity:

\[
\pi_k \propto \exp \left( - \frac{||v_{j^*} - u_k||^2_2}{2(1-\delta)} \right)
\]

With this prior, the entropy becomes:

\[
H(\pi) = \sum_k \pi_k \log \left( \frac{\sum_i \exp \left( - \frac{||v_{j^*} - u_k||^2_2}{2(1-\delta)} \right)}{\exp \left( - \frac{||v_{j^*} - u_k||^2_2}{2(1-\delta)} \right)} \right)
\]

\[
= \log(W(\mathcal{V}, 2(1-\delta))) + 1 + \sum_k \pi_k \frac{||v_{j^*} - u_k||^2_2}{2(1-\delta)}
\]

\[
= \log(W(\mathcal{V}, 2(1-\delta))) + 1 + \frac{1}{1-\delta} \sum_k \pi_k KL(P_k || P_{j^*})
\]
The 1 inside the first log comes from the fact that in the definition $W_{j^*}$, we do not include the term involving $j^*$ in the sum, while our prior $\pi$ does place mass proportional to 1 on hypothesis $j^*$. The term involving the KL-divergence follows from the fact that the KL between two gaussians is one-half the $\ell_2^2$-distance between their means. 

Looking at the lower bound from Fano’s inequality, we see that if:

$$E_k \sim \pi KL(P_k || P_\pi) + \log 2 \leq (1 - \delta) H(\pi) = (1 - \delta) \log(W(V, 2(1 - \delta)) + 1) + \sum_k \pi_k KL(P_k || P_{j^*})$$

then the probability of error is lower bounded by $\delta$. Of course it is immediate that:

$$\sum_k \pi_k KL(P_k || P_{j^*}) = \sum_k \pi_k \int P_k(x) \log \frac{P_k(x) P_\pi(x)}{P_\pi(x) P_{j^*}(x)} = \sum_k \pi_k KL(P_k || P_\pi) + KL(P_\pi || P_{j^*}) \geq E_k KL(P_k || P_\pi)$$

So the condition reduces to requiring that:

$$\log 2 \leq (1 - \delta) \log(W(V, 2(\delta - 1) + 1).$$

After some algebra, this is equivalent to:

$$W(V, 2(\delta - 1)) \geq 2^{\frac{1}{1-\delta}} - 1$$

\[\square\]

### A.2 Proof of Theorem 7

The proof of Theorem 7 is essentially the same as the proof of Theorem 2, coupled with two observations. First, for a sampling strategy $B \in \mathbb{R}^d_+$ the maximum likelihood estimator is:

$$T_{\text{MLE}}(y, B) = \arg\min_{j \in [M]} \| v_j - y \|_B^2$$

so the analysis of the MLE depends on the Mahalanobis norm $\| \cdot \|_B$ instead of the $\ell_2$ norm.

Similarly, the KL divergence between the distribution $P_{j, B}$ and $P_{k, B}$ depends on the Mahalanobis norm $\| \cdot \|_B$ instead of the $\ell_2$ norm. Specifically, we have:

$$KL(P_{j, B} || P_{k, B}) = \frac{1}{2} \| v_j - v_k \|_B^2.$$ 

The lower bound proof instead use this metric structure, but the calculations are equivalent. \[\square\]

### A.3 Proof of Proposition 8

To simplify the presentation, let $f(B) = W(V, \alpha, B)$. $f(B)$ is convex and (strictly) monotonically decreasing, so we know that the minimum will be achieved when the constraint is tight, i.e. when $\|B\|_1 = \tau$. The Lagrangian is:

$$\mathcal{L}(B, \lambda) = f(B) + \lambda(\|B\|_1 - \tau)$$

15
and the minimum is achieved at $\hat{B}$, with $\|\hat{B}\|_1 = \tau$, if there is a value $\lambda$ such that $0 \in \partial L(\hat{B}, \hat{\lambda})$. Observing that the subgradient is $\partial f(B) + \lambda 1$, it suffices to ignore the Lagrangian term and instead ensure that $\|\partial f(B)\| < 1$. $f(B)$ is a maximum of $M$ convex functions, where $f_j(B)$ is the function corresponding to hypothesis $v_j$, and, by direct calculation, the subgradient of this function $f_j(B)$ is:

$$\frac{\partial f_j(B)}{\partial B_i} = \sum_{k \neq j} -(v_k(i) - v_j(i))^2 \exp(-\|v_k - v_j\|_2^2).$$

Moreover, the subgradient of the maximum of a set of functions is the convex hull of the subgradients of all functions achieving the maximum. This means that if there exists a distribution $\pi$, supported over the maximizers of $f(B)$, such that the expectation of the subgradients is constant, we have certified optimality of $\hat{B}$. This is precisely the condition in the Proposition. \hfill \square

### A.4 Proof of Theorem 6

**Proof of Proposition 3** To prove Proposition 3, we make two claims. First we certify that for a prior $\pi$, the Maximum a Posteriori (MAP) estimator is a Bayes estimator for prior $\pi$. Given $\pi$, the map estimator is:

$$T_\pi(y) = \arg\max_j \pi_j \exp\{-\|v_j - y\|_2^2/2\}$$

Define the posterior risk of an estimator $T$ to be the expectation of the loss, under the posterior distribution on the hypothesis. In our case this is:

$$r(T|y) = \sum_{j=1}^{M} 1[T(y) \neq j] \pi_j(y) \text{ where } \pi_j(y) \propto \pi_j \exp\{-\|v_j - y\|_2^2/2\}.$$ 

For a fixed $y$, this quantity is minimized by letting $T(y)$ be the maximizer of the posterior, as this makes the $0 - 1$ loss term zero for the largest $\pi_j(y)$ value. Thus for each $y$ we minimize the posterior risk by letting $T(y)$ be the MAP estimate. The result follows by the well known fact that if an estimator minimizes the posterior risk at each point, then it is the Bayes estimator.

This argument shows that the only types of estimators we need to analyze are MAP estimators under various priors. This gives us the requisite structure to prove Proposition 3.

Specifically, for a prior $\pi$, for the MAP estimate to predict hypothesis $j$, it must be the case that:

$$\forall k \neq j. \quad \pi_j \exp\{-\|v_j - y\|_2^2/2\} \geq \pi_k \exp\{-\|v_j - v_k\|_2^2/2\}.$$ 

This can be simplified to:

$$\langle v_j - v_k, y \rangle \geq \frac{1}{2}(\|v_j\|_2^2 - \|v_k\|_2^2) + \log \frac{\pi_k}{\pi_j}.$$ 

Thus the acceptance region for the hypothesis $j$ is the set of all points $y$ that satisfy all of these $M - 1$ inequalities. This is exactly the polyhedral set $A_j$. \hfill \square

**Proof of Proposition 4** We provide a proof of this well-known result showing that the Bayes estimator with uniform risk landscape is minimax optimal. Let $\hat{T}_\pi$ be the Bayes estimator under prior $\pi$ and let $T_0$ be some other estimator. Since $\hat{T}_\pi$ has constant risk landscape, we know that $\max_j R_j(V, \hat{T}_\pi) = B_\pi(\hat{T}_\pi)$, or the minimax risk for $\hat{T}_\pi$ is equal to its Bayes risk. We know that the Bayes risk of $T_0$ is at most the minimax
risk for $T_0$, i.e. $B_\pi(T_0) \leq \max_j \mathcal{R}_j(\mathcal{V}, T_0)$. If it were the case that $T_0$ had strictly lower minimax risk, then we have:

$$B_\pi(T_0) \leq \max_j \mathcal{R}_j(\mathcal{V}, T_0) < \max_j \mathcal{R}_j(\mathcal{V}, T_\pi) \leq B_\pi(T_\pi).$$

However, this is a contradiction since $T_\pi$ is the Bayes estimator under prior $\pi$, meaning that it minimizes the Bayes risk.

Proof of Theorem 6: Our goal is to apply Proposition 4. By the fact that $\mathcal{R}_j(\mathcal{V}, T) = 1 - \mathbb{P}_j[A_j]$ where $A_j$ is $T$’s acceptance region for hypothesis $j$, we must show that the $\mathbb{P}_j$ probability content of the acceptance regions are constant. Ignoring the normalization factor of the gaussian density, this is:

$$\int_{A_j} \exp\left\{-||v_j - x||^2/2\right\} dx,$$

where $A_j = \{z | \Gamma_j z \geq b_j\}$ as defined in Proposition 3. We will exploit the unitary invariance of the family.

For any pair of hypothesis $j, k$, let $R_{jk}$ be the orthogonal matrix such that $v_k = R_{jk} v_j$ and note that $R_{kj}$, the orthogonal matrix that maps $v_j$ to $v_j$, is just $R_{jk}^T$. This also means that $R_{jk} R_{jk}^T = R_{jk} R_{kj} = I$. Via a change of variables $x = R_{kj} y$, the integrand becomes:

$$\exp\left\{ -||v_j - R_{kj} y||^2/2 \right\} = \exp\left\{ -||R_{jk} v_j - R_{jk} R_{kj} y||^2/2 \right\} = \exp\left\{ -||v_j - y||^2/2 \right\}.$$

Thus, we have translated to the $P_k$ measure.

As for the region of integration, first note that since $v_i = R_{jk} v_j$, it must be the case that $||v_j||^2_2 = ||v_i||^2_2$ for all $j, i \in [M]$. This means that the vector $b_j$ defining the acceptance region, which for the MLE has coordinates $b_j(i) = \frac{1}{2}(||v_j||^2_2 - ||v_i||^2_2)$, is just the all-zeros vector. The region of integration is therefore:

$$\{z | \Gamma_j z \geq 0\} = \{z | \Gamma_j R_{kj} z \geq 0\}.$$

We must check that this polytope is exactly $A_k$, which means that we must check that for each $i$, $(v_j - v_i)^T R_{kj}$ is a row of the $\Gamma_k$ matrix. But:

$$(v_j - v_i)^T R_{kj} = v_j^T R_{jk}^T - v_i^T R_{jk}^T = v_k^T - v_i^T R_{jk}^T.$$

Since $v_i$ can generate the family $\mathcal{V}$, it must be the case that $R_{jk} v_i \in \mathcal{V}$ so that this difference does correspond to some row of $\Gamma_k$. Since we apply the same unitary operator to all of the rows, it must be the case that the number of distinct rows is unchanged, or in other words, there is a bijection from the rows in $\Gamma_j R_{kj}$ to the rows in $\Gamma_k$. Therefore, the transformed region of integration, after the change of variable $x = R_{kj} y$, is exactly the acceptance region $A_k$, and the integrand is the $P_k$ measure. This means that $P_k[A_k] = P_j[A_j]$ and this is true for all pairs $(j, k)$, so that the risk landscape is constant. By Proposition 4 this certifies optimality of the MLE.

A.5 Calculations for the examples

Calculations for $k$-Sets: We must upper and lower bound $W(\mathcal{V}, \alpha)$. First note that by symmetry, every hypothesis achieves the maximum, so it suffices to compute just one of them:

$$W(\mathcal{V}, \alpha) = \sum_{k \neq j} \exp\left\{-||v_k - v_j||^2_2/\alpha\right\} = \sum_{s=1}^{k} \binom{k}{s} \left(\frac{d-k}{s}\right) \exp\left(-2s\mu^2/\alpha\right).$$
This follows by noting that the $\ell_2^2$ distance between two hypothesis is the symmetric set difference between the two subsets, and then by a simple counting argument. Using well known bounds on binomial coefficients, we obtain:

$$W(V, \alpha) \leq \sum_{s=1}^{k} \exp(s \log(ke/s) + s \log((d - k)e/s) - 2s\mu^2/\alpha)$$

$$= k \sum_{s=1}^{k} \exp(s \log(e^2k(d - k)/s^2) - 2s\mu^2/\alpha)$$

$$\leq k \exp(\log e^2k(d - k) - 2\mu^2/\alpha) \quad \text{if} \ 2\mu^2/\alpha \geq \log(e^2k(d - k))$$

This is smaller than $\delta$ whenever $\mu^2 \geq \alpha \log(ek(d - k)/\delta)$, which subsumes the requirement above. For the lower bound:

$$W(V, \alpha) \geq \sum_{s=1}^{k} \exp(s \log(k/s) + s \log((d - k)/s) - 2s\mu^2/\alpha) \geq \exp(-2\mu^2/\alpha + \log(k(d - k)))$$

which goes to infinity if $\mu^2 = o(\alpha \log(k(d - k)))$.

To certify that the uniform allocation strategy minimizes $W(V, \alpha, B)$, we apply Proposition 8. Fix $\tau$ and let $\hat{B}$ be such that $\hat{B}(i) = \tau/d$. By symmetry, every hypothesis achieves the maximum under this allocation strategy, and we will take $\pi$ to be the uniform distribution over all hypothesis.

For a hypothesis $j$ and a coordinate $i$, the subgradient $\partial f_j(B) / \partial B(i)$ at $\hat{B}$ depends on the whether $v_j(i) = 0$ or not. If $v_j(i) = 0$, then:

$$\frac{\partial f_j(B)}{\partial B(i)} = \mu^2 \sum_{s=1}^{k} \binom{d - k - 1}{s - 1} \binom{k}{s} \exp(-2\mu^2 s^2/d),$$

and if $v_j(i) = \mu^2$ then:

$$\frac{\partial f_j(B)}{\partial B(i)} = \mu^2 \sum_{s=1}^{k} \binom{d - k}{s} \binom{k - 1}{s - 1} \exp(-2\mu^2 s^2/d).$$

Both of these follow from straightforward counting arguments. Notice that the value of the subgradient depends only on whether $v_j(i) = 0$ or not, and under the uniform distribution $\pi$, $E_j \sim \pi v_j(i) = E_j \sim \pi v_j(i')$. This implies that the constant vector is in the subgradient of $f(B)$ at $\hat{B}$, so that $\hat{B}$ is the minimizer of $W(V, \alpha, B)$ subject to $\|B\|_1 \leq \tau$.

We have already done the requisite calculation to bound the minimax risk under sampling. The calculations above show that if $\mu = \omega(\sqrt{\frac{d}{\tau} \log(k(d - k))})$ then the maximum likelihood estimator, when using the uniform sampling strategy has risk tending to zero. Conversely if $\mu = o(\sqrt{\frac{d}{\tau} \log(k(d - k))})$ then the minimax risk, for any allocation strategy tends to one.

**Calculation for Biclusters:** Due to symmetry, all hypotheses achieve the maximum and therefore, we
These bounds lead to the same minimax rate as above. The two terms here are identical, so we will just bound the first one:

\[
W(\mathcal{V}, \alpha) = \sum_{s_r=1}^{k} \sum_{s_c=1}^{k} C_k^r C_k^c C_{d-k}^s C_{d-k}^c \exp \left( -\frac{2\mu^2}{\alpha} (s_r(k - s_c) + s_c(k - s_r) + s_r s_c) \right) \\
+ \sum_{s_r=1}^{k} C_k^r C_{d-k}^c \exp \left( -\frac{2\mu^2}{\alpha} (s_r k) \right) + \sum_{s_c=1}^{k} C_k^c C_{d-k}^r \exp \left( -\frac{2\mu^2}{\alpha} (s_c k) \right)
\]

This last two term comes from the case where \( s_c = 0 \) or \( s_r = 0 \), which is all of the hypotheses that share the same columns but disagree on the rows (or share the same rows but disagree on the columns). Using binomial approximations, the first term can be upper bounded by:

\[
\leq \sum_{s_r=1}^{k} \sum_{s_c=1}^{k} \exp \left( s_r \log \frac{k(d-k)\mu^2}{s_r^2} + s_c \log \frac{k(d-k)\mu^2}{s_c^2} - \frac{2\mu^2}{\alpha} (s_r (k - s_c/2) + s_c (k - s_r/2)) \right) \\
\leq \sum_{s_r=1}^{k} \exp \left( s_r \left( \log \frac{k(d-k)\mu^2}{s_r^2} - \frac{k\mu^2}{\alpha} \right) \right) \sum_{s_c=1}^{k} \exp \left( s_c \left( \log \frac{k(d-k)\mu^2}{s_c^2} - \frac{k\mu^2}{\alpha} \right) \right).
\]

The two terms here are identical, so we will just bound the first one:

\[
\sum_{s_r=1}^{k} \exp \left( s_r \left( \log \frac{k(d-k)\mu^2}{s_r^2} - \frac{k\mu^2}{\alpha} \right) \right) \\
\leq \sum_{s_r=1}^{k} \exp \left( s_r \left( \log (k(d-k)\mu^2) - \frac{k\mu^2}{\alpha} \right) \right) \\
\leq k \exp \left( \log (k(d-k)\mu^2) - \frac{k\mu^2}{\alpha} \right) \quad \text{if } \mu^2 \geq \frac{\alpha}{k} \log (k(d-k)\mu^2)
\]

Applying this inequality to both terms gives a bound on \( W(\mathcal{V}, \alpha) \). This bound is smaller than \( \delta \) as long as \( \mu \geq \sqrt{\frac{\log k(d-k)\mu^2}{\alpha}} \) for some universal constant \( c \). Again this subsumes the condition required for the inequality to hold.

The other two terms are essentially the same. Using binomial approximations, both expressions can be bounded as:

\[
\sum_{s_c=1}^{k} C_k^c C_{d-k}^r \exp \left( -\frac{2\mu^2}{\alpha} (s_c k) \right) = \sum_{s_r=1}^{k} \exp \left( s_r \log (e^2k(d-k)/s_r^2) - 2s_r k \mu^2 / \alpha \right) \\
\leq k \exp \left( \log (k(d-k)\mu^2) - 2k \mu^2 / \alpha \right) \quad \text{if } \mu^2 \geq \frac{\alpha}{2k} \log (k(d-k)\mu^2).
\]

These bounds lead to the same minimax rate as above.
For the lower bound, we again use binomial approximations.

\[
W(V, \alpha) \geq \sum_{s_r=1}^{k} \sum_{s_c=1}^{k} \exp \left( s_r \log \frac{k(d-k)}{s_r^2} + s_c \log \frac{k(d-k)}{s_c^2} - \frac{2\mu^2}{\alpha} (s_r(k-s_c/2) + s_c(k-s_r/2)) \right)
\]

\[
\geq \sum_{s_r=1}^{k} \exp \left( s_r \left( \log \frac{k(d-k)e^2}{s_r^2} - \frac{2k\mu^2}{\alpha} \right) \right) \sum_{s_c=1}^{k} \exp \left( s_c \left( \log \frac{k(d-k)e^2}{s_c^2} - \frac{2k\mu^2}{\alpha} \right) \right)
\]

\[
\geq \exp(\log(k(d-k)) - 2\mu^2k/\alpha)^2
\]

This lower bound goes to infinity if \( \mu = o(\sqrt{\frac{1}{k} \log(k(d-k))}) \) lower bounds the minimax rate.

To certify that the uniform allocation strategy minimizes \( W(V, \alpha, B) \), we apply Proposition. Fix \( \tau \) and let \( \bar{B} \) be such that \( \bar{B}(a,b) = \tau/d^2 \) for all \( (a,b) \in [d] \times [d] \). By symmetry, every hypothesis achieves the maximum under this allocation strategy, and we will take \( \pi \) to be the uniform distribution over all hypothesis.

For a hypothesis \( j \), let \( f_j(B) \) denote the term in the SEDF centered around \( j \). For a hypothesis \( j \) based on clusters \( S_1, S_r \) and a coordinate \( (a,b) \), the subgradient \( \frac{\partial f_j(B)}{\partial B(a,b)} \) at \( \bar{B}(a,b) \) depends on whether \( a \in S_1 \) and \( b \in S_r \). If \( a \notin C_l \) and \( b \notin C_r \), then:

\[
\frac{\partial f_j(B)}{\partial B(a,b)} \bigg|_{B=\bar{B}} = \frac{-\mu^2}{\alpha} \sum_{s_r=1}^{k} \sum_{s_c=1}^{k} C^{sr}_{d-k-1} C^{sc}_{d-k-1} C^{rk}_{k} \exp \left( \frac{-2\mu^2}{\alpha d^2} (s_r(k-s_c/2) + s_c(k-s_r/2)) \right).
\]

This follows by direct calculation. Similar calculations yield the other cases:

\[
\frac{\partial f_j(B)}{\partial B(a,b)} \bigg|_{B=\bar{B}} = \frac{-\mu^2}{\alpha} \sum_{s_r=0}^{k-1} \sum_{s_c=1}^{k} C^{sr}_{d-k-1} C^{sc}_{d-k-1} C^{rk}_{k} \exp \left( \frac{-2\mu^2}{\alpha d^2} (s_r(k-s_c) + s_c(k-s_r) + s_r s_c) \right).
\]

\[
\frac{\partial f_j(B)}{\partial B(a,b)} \bigg|_{B=\bar{B}} = \frac{-\mu^2}{\alpha} \sum_{s_r=1}^{k} \sum_{s_c=0}^{k} C^{sr}_{d-k-1} C^{sc}_{d-k-1} C^{rk}_{k} \exp \left( \frac{-2\mu^2}{\alpha d^2} (s_r(k-s_c) + s_c(k-s_r) + s_r s_c) \right).
\]

\[
\frac{\partial f_j(B)}{\partial B(a,b)} \bigg|_{B=\bar{B}} = \frac{-\mu^2}{\alpha} \sum_{s_r=0}^{k} \sum_{s_c=0}^{k} C^{sr}_{d-k-1} C^{sc}_{d-k-1} C^{rk}_{k} \exp \left( \frac{-2\mu^2}{\alpha d^2} (s_r(k-s_c) + s_c(k-s_r) + s_r s_c) \right).
\]

These correspond to the cases \( a \in S_1, b \notin S_r, a \notin S_1, b \in S_r \) and the case where \( a \in S_1, b \in S_r \) respectively. The main point is that the value of the subgradient depends only on presence or absence of the row/column in the cluster, and under the uniform distribution \( \pi \), each row/column is equally likely to be in the cluster. This means that for every coordinate \( (a,b) \) taking the expected subgradient with respect to the uniform distribution over hypotheses yields the same expression. So the constant vector is in the subgradient of \( f(B) \) at \( \bar{B} \), so that \( \bar{B} \) is the minimizer of \( W(V, \alpha, B) \) subject to \( \|B\|_1 \leq \tau \).

We have already done the requisite calculation to bound the minimax risk under sampling. The calculations above show that if \( \mu = \omega(\sqrt{\frac{1}{k} \log(k(d-k))}) \) then the maximum likelihood estimator, when using the uniform sampling strategy has risk tending to zero. Conversely if \( \mu = o(\sqrt{\frac{1}{k} \log(k(d-k))}) \) then the minimax risk, for any allocation strategy tends to one.

The biclusters family is clearly unitarily invariant with respect to the set of orthonormal matrices that permute the rows and columns independently. The family is easiest to describe as acting on the matrices \( 1_{S_1}1_{S_r}^T \). Let \( P_l, P_r \) be any two \( d \times d \) permutation matrices. Then the matrix \( P_l 1_{S_1} (1_{S_r} P_r)^T \) is clearly another hypothesis, and as we vary \( P_l \) and \( P_r \) we generate all of the hypothesis. Note that these permutations are unitary operators on the matrix space \( \mathbb{R}^{d \times d} \), which allows us to apply Theorem.
For the analysis of the interactive algorithm, let us first bound the probability that the algorithm makes a mistake on any single coordinate. Consider sampling a coordinate \( x \) with mean \( \mu \) and noise variance \( 1/b \). A Gaussian tail bound reveals that:

\[
P[|x - \mu| \geq \epsilon] \leq 2 \exp(-2b\epsilon^2).
\]

We will sample no more than \( d^2 \) coordinates and we will sample each coordinate with the same amount of energy \( b \). So by the union bound, the probability that we make a single mistake in classifying a coordinate that we query is bounded by \( \delta/2 \) as long as:

\[
\mu \geq 2\epsilon = \sqrt{\frac{2}{b}} \log(4d^2/\delta).
\]

We now need to bound \( b \), which depends on the total number of coordinates queried by the algorithm. In the first phase of the algorithm, we sample coordinates uniformly at random until we hit one that is active. Since each sample hits an active coordinate with probability \( \mu \) and plugging into the condition on \( \mu \) above proves the result.

**Calculation for Stars:** For the stars problem, define \( \mathcal{N}(j) \subseteq V \) to be the neighbors of the vertex \( j \) in the graph. For a fixed hypothesis \( j \), we have

\[
W_j(\mathcal{V}, \alpha) = \sum_{k \neq j} \exp \left( -\|v_k - v_j\|_2^2 / \alpha \right)
\]

\[
= \sum_{k \in \mathcal{N}(j)} \exp(-\mu^2(\deg(k) + \deg(j) - 2) / \alpha) + \sum_{k \notin \mathcal{N}(j)} \exp(-\mu^2(\deg(k) + \deg(j)) / \alpha)
\]

\[
\leq \exp\left( -\mu^2\deg_{\min} / \alpha - \mu^2\deg(j) / \alpha \right) \left( \deg(j) \exp(2\mu^2 / \alpha) + |V| - \deg(j) \right)
\]

This last inequality follows by replacing every \( \deg(k) \) with \( \deg_{\min} \), the lower bound on the degrees. This last expression is maximized with \( \deg(j) = \deg_{\min} \), which can be observed by noticing that the derivative with respect to \( \deg(j) \) is negative. This gives the bound:

\[
W(\mathcal{V}, \alpha) \leq \exp\left( -2\mu^2\deg_{\min} / \alpha \right) \left( \deg_{\min} \exp(2\mu^2 / \alpha) + |V| - \deg_{\min} \right)
\]

One can lower bound \( W(\mathcal{V}, \alpha) \) by choosing the hypothesis \( j \) with \( \deg(j) = \deg_{\min} \) and then replacing all other degree terms with \( \deg_{\min} \) in the above calculations. This gives:

\[
W(\mathcal{V}, \alpha) \geq \exp\left( -\mu^2 \deg_{\min} / \alpha \right) \left( \deg_{\min}^{2\mu^2 / \alpha} + |V| - \deg_{\min} \right)
\]

**Calculation for Random Codes:** In the proof of Theorem 2, we saw that for a hypothesis \( j \), we can bound the probability of error by:

\[
\mathbb{P}_j[\text{error}] \leq \sum_{k \neq j} \exp(-\|v_j - v_k\|_2^2 / 8)
\]

21
This means that:

\[
\mathbb{E}_V \sum_{j=1}^M \mathbb{P}_j \left[ \text{error} \right] \leq \mathbb{E}_V \frac{1}{M} \sum_{j=1}^M \sum_{k \neq j} \exp(-\|v_j - v_k\|^2/8)
\]

\[
= (M - 1) \mathbb{E}_{v,v'} \exp(-\|v - v\'|^2/8) = (M - 1) \prod_{j=1}^d \mathbb{E}_{x \sim \chi^2_1} \exp(-Px/4)
\]

\[
= (M - 1)(1 + P/2)^{-d/2}.
\]

Notice that the only inequality in this sequence is the first one, which is essentially an application of Theorem 2. The last equality is based on the moment-generating function of a \( \chi^2_1 \) random variable.

To achieve the bound on the rate of the code, set this final expression to be at most some \( f(d) \) which is \( o(1) \). Then the probability of error is at most \( f(d) \to 0 \) and the rate \( R \) is:

\[
R = \frac{\log M}{d} = \frac{1}{2} \log(1 + P/2) + \frac{\log(f(d))}{d} = \frac{1}{2} \log(1 + P/2) - \omega(1/d)
\]