entropy

MDPI

Review

On Lower Bounds for Statistical Learning Theory

Po-Ling Loh

Department of Electrical and Computer Engineering, University of Wisconsin-Madison, 1415 Engineering Drive, Madison, WI 53706, USA; loh@ece.wisc.edu; Tel.: +1-443-968-5029

Received: 7 September 2017; Accepted: 14 November 2017; Published: 15 November 2017

Abstract: In recent years, tools from information theory have played an increasingly prevalent role in statistical machine learning. In addition to developing efficient, computationally feasible algorithms for analyzing complex datasets, it is of theoretical importance to determine whether such algorithms are "optimal" in the sense that no other algorithm can lead to smaller statistical error. This paper provides a survey of various techniques used to derive information-theoretic lower bounds for estimation and learning. We focus on the settings of parameter and function estimation, community recovery, and online learning for multi-armed bandits. A common theme is that lower bounds are established by relating the statistical learning problem to a channel decoding problem, for which lower bounds may be derived involving information-theoretic quantities such as the mutual information, total variation distance, and Kullback–Leibler divergence. We close by discussing the use of information-theoretic quantities to measure independence in machine learning applications ranging from causality to medical imaging, and mention techniques for estimating these quantities efficiently in a data-driven manner.

Keywords: machine learning; minimax estimation; community recovery; online learning; multi-armed bandits; channel decoding; threshold phenomena

1. Introduction

Statistical learning theory refers to the rigorous mathematical analysis of machine learning algorithms [1,2]. On one hand, it is desirable to derive error bounds for the performance of particular machine learning algorithms under appropriate assumptions on the probabilistic models used to generate the data. On the other hand, it is important to understand the fundamental limitations of any algorithmic procedure, which may be influenced by quantities such as the sample size, signal-to-noise ratio, or smoothness of an ambient function space. Whereas statistical techniques based on concentration inequalities and empirical process theory may often be employed to derive rates of convergence of specific estimators to the underlying parameters of a data-generating distribution, the somewhat trickier problem of quantifying the best possible performance of *any* learning procedure requires tools from information theory.

A general approach is to relate the machine learning task at hand to an appropriate channel decoding problem, where the output corresponds to the observed data and the input corresponds to a cleverly constructed subset of the parameter space. For estimation problems, the key observation is that, if the underlying parameters may be estimated closely (i.e., on the level of discretization of the subset of parameter space), decoding may be performed accurately with high probability. The hardness of the decoding problem may in turn be quantified using techniques in information theory [3], leading to a lower bound on the estimation error. This strategy has been applied successfully to a diverse array of statistical estimation problems, including parametric and nonparametric regression, structure estimation for graphical models, covariance matrix estimation, and dimension reduction methods such as principal component analysis [4–9]. Section 2 discusses the method and several illustrative examples in greater detail.

Although some classes of machine learning problems may not be analyzed directly using these methods, alternative approaches involving related information-theoretic concepts may be employed. In Sections 3 and 4, we consider the problems of community recovery and online learning, which are both active areas of research in machine learning. Our discussion of weak recovery in the community estimation setting is similar to the framework described in Section 2, but since the loss function used to quantify the estimation error incurred by the algorithm is more complicated, a more careful analysis must be conducted to derive sharp lower bounds. The theory characterizing the regimes in which exact recovery is possible are of a somewhat different flavor, but the emergence of sharp thresholds may again be related to Shannon coding theory. Section 4, concerning online learning for multi-armed bandits, provides a still different setting, where the goal is to bound a quantity known as regret. Although this is a radically different goal from bounding estimation error, the techniques used to obtain lower bounds for multi-armed bandits nonetheless include components of reductions to channel decoding problems: The key is to relate the performance of a learning algorithm to a problem of distinguishing between pairs of parameter assignments corresponding to underlying reward distributions that are close in parameter space.

We include proof sketches for the stated theorems in the main text of the paper, with references to resources where the reader can find more detailed proofs and additional background material. Although the discussion of each problem setting is necessarily brief, given the broad scope of this paper, we hope that our survey will convey the high-level ideas involved in applying information-theoretic tools to derive lower bounds for some statistical machine learning problems in a clear, concise manner. We have intentionally selected a diverse variety of problem settings in order to help the reader compare and contrast different approaches for obtaining lower bounds and identify the common threads underlying all the strategies.

2. Statistical Estimation

We begin by discussing an approach based on minimax theory for statistical estimation problems [10]. Our goal is a lower bound on the following quantity, known as the *minimax risk*:

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{X \sim P}[\ell(\widehat{\theta}(X), \theta(P))], \tag{1}$$

where ℓ is a symmetric loss function. Here, $\mathcal{P}$ denotes a class of data-generating distributions and $\theta : \mathcal{P} \to \Omega$ is a functional that maps each distribution in $\mathcal{P}$ to a parameter in the metric space Ω. The expectation in expression (1) is taken with respect to data from a particular distribution $P \in \mathcal{P}$, and the infimum is then taken over all possible estimators $\widehat{\theta} = \widehat{\theta}(X)$ computed from the data. In other words, quantity (1) captures the worst-case risk of the best possible estimator. Whereas statistical analysis of a specific estimator can provide an upper bound on the minimax risk, tools from information theory may be used to derive a lower bound on the same quantity. Throughout this section, we will restrict our attention to the setting where $\ell = \Phi \circ \rho$, for a metric ρ and monotonically increasing function $\Phi : [0, \infty) \to [0, \infty)$. For instance, Example 2 below will discuss the setting where ρ is the L_2-distance in a function space and $\Phi(t) = t^2$, so ℓ is the squared L_2-distance.

The basic idea is to transform an estimation problem into a decoding problem, in which we wish to infer the correct message from a discrete set of messages, corresponding to a collection of parameters. The estimation problem must be at least as hard as the decoding problem, since, if the parameters in the discrete set are appropriately separated, accurate parameter estimation implies accurate decoding. In Section 2.1, we present a general technique based on Fano's inequality, which expresses the probability of error for the decoding in terms of the mutual information between the input (parameters in the discrete subset) and output (observed data). Sections 2.2 and 2.3 then provide methods for bounding the mutual information and discuss applications to concrete statistical estimation settings. We will follow the convention of Cover and Thomas [3] and take all logarithms with respect to base 2 in our definitions of entropy and mutual information; analogous results hold when logarithms are taken with respect to base e.

2.1. Fano's Method

We begin by describing the general approach for deriving lower bounds. The key idea consists of relating the estimation problem to a decoding problem, and then using Fano's inequality to lower-bound the probability of error for the decoding problem. Recall the definition of the mutual information:

$$I(Y;X) = H(Y) - H(Y|X). \tag{2}$$

The main result relates the minimax risk to the mutual information between observations and the data-generating distribution.

Theorem 1. *Suppose* $\{P_1, \ldots, P_M\} \subseteq \mathcal{P}$ *satisfy* $\rho(\theta(P_i), \theta(P_j)) \geq 2\delta$, *for all* $i \neq j$. *Then*

$$\inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_{X \sim P}[\ell(\widehat{\theta}(X), \theta(P))] \geq \Phi(\delta) \left(1 - \frac{I(Y;X) - 1}{\log_2 M}\right),$$

where Y *is distributed uniformly on* $\{1, \ldots, M\}$ *and the conditional distribution of* X *given* Y *is defined by* $X \mid \{Y = j\} \sim P_j$.

Proof (sketch). We begin by writing

$$\sup_{P \in \mathcal{P}} \mathbb{E}_{X \sim P}[\ell(\widehat{\theta}(X), \theta(P))] \geq \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{X \sim P_i}[\ell(\widehat{\theta}(X), \theta(P_i))]. \tag{3}$$

If we define the decision rule

$$\psi(X) = \arg \min_{1 \leq j \leq M} \ell(\theta_j, \widehat{\theta}(X)),$$

where we break ties arbitrarily, we may verify that

$$\mathbb{E}_{X \sim P_i}[\ell(\widehat{\theta}(X), \theta(P_i))] \overset{(a)}{\geq} \Phi(\delta) \mathbb{P}_i \left(\ell(\widehat{\theta}(X), \theta(P_i)) \geq \Phi(\delta)\right) \overset{(b)}{\geq} \Phi(\delta) \mathbb{P}_i \left(\psi(X) \neq i\right),$$

for each $1 \leq i \leq M$. Inequality (a) is a direct application of Markov's inequality, and inequality (b) follows from the fact that if $\ell(\widehat{\theta}, \theta_i) < \Phi(\delta)$, or equivalently, $\rho(\widehat{\theta}, \theta_i) < \delta$, then

$$\rho(\widehat{\theta}, \theta_j) \geq \rho(\theta_i, \theta_j) - \rho(\theta_i, \widehat{\theta}) > 2\delta - \delta > \rho(\widehat{\theta}, \theta_i), \qquad \forall j \neq i,$$

implying that $\psi(X) = i$.

Now, recall the statement of Fano's inequality:

Lemma 1 (Fano's inequality [3])**.** *For any estimator* $\widehat{Y}$ *of* Y *such that* $Y \to X \to \widehat{Y}$ *forms a Markov chain, it holds that*

$$\mathbb{P}(\widehat{Y} \neq Y) \geq \frac{H(Y|X) - 1}{\log_2 |\mathcal{Y}|},$$

where $\mathcal{Y}$ *is the range of* Y.

Applying Lemma 1 with $\widehat{Y} = \psi(X)$ and writing out the error probability explicitly, we obtain

$$\frac{1}{M} \sum_{i=1}^{M} \mathbb{P}_i(\psi(X) \neq i) \geq \frac{H(Y|X) - 1}{\log_2 M} = \frac{\log_2 M - I(Y;X) - 1}{\log_2 M}, \tag{4}$$

where the equality follows from relation (2) and the fact that Y has a uniform distribution. Combining inequalities (3) and (4) establishes the desired result. □

In the following subsections, we describe two methods for upper-bounding the mutual information term $I(Y;X)$ appearing in Theorem 1, yielding a lower bound on the minimax risk.

2.2. Local Packings

The first method applies the convexity of the Kullback-Leibler (KL) divergence to obtain an upper bound on $I(Y;X)$ in terms of pairwise KL divergences. We have the following lemma:

Lemma 2. *Let X and Y be defined as in Theorem 1. Then,*

$$I(Y;X) \leq \frac{1}{M^2} \sum_{1 \leq i,j \leq M} D_{KL}(P_i \| P_j).$$

Proof. We can check that

$$I(Y;X) = \frac{1}{M} \sum_{i=1}^{M} D_{KL}(P_i \| \overline{P}),$$

where $\overline{P} = \frac{1}{M} \sum_{j=1}^{M} P_j$ is a mixture distribution. By the convexity of the KL divergence, we then have

$$I(Y;X) \leq \frac{1}{M} \sum_{i=1}^{M} \frac{1}{M} \sum_{j=1}^{M} D_{KL}(P_i \| P_j),$$

which is the desired expression. □

This bounding technique is known as a "local packing", since the trick is to design an appropriate set $\{P_1, \ldots, P_M\}$ such that the parameters $\theta(P_i)$ are 2δ-separated, while the pairwise KL divergences between the data-generating distributions are relatively small.

Example 1 (High-dimensional linear regression)**.** *Suppose we have observation pairs $\{(x_i, y_i)\}_{i=1}^n$ from a linear model:*

$$y_i = x_i^T \beta^* + w_i,$$

where $x_i \in \mathbb{R}^p$ and $w_i \sim N(0, \sigma^2)$ is i.i.d. noise, and $\beta^* \in \mathbb{R}^p$ is the unknown parameter vector. We assume that $p > n$, but β^* is known to have at most s nonzero values, where $s \leq n$. More precisely, if $\mathbb{B}_q(r)$ denotes the ball of radius r in the ℓ_q norm, we are interested in characterizing the minimax risk over the parameter space

$$\mathbb{B}_0(s) \cap \mathbb{B}_2(1) = \{\beta \in \mathbb{R}^p : \|\beta\|_0 \leq s, \|\beta\|_2 \leq 1\}.$$

For any fixed parameter $\delta > 0$, it is possible to construct a subset of parameters $\{\beta_1, \ldots, \beta_M\}$ lying in the parameter space such that $\delta \leq \|\beta_j - \beta_k\|_2 \leq 2\delta\sqrt{2}$ for all $1 \leq j < k \leq M$ and $\log M \geq \frac{s}{2} \log\left(\frac{p-s}{s/2}\right)$, essentially by rescaling a packing of the subset of $\{-1, 0, 1\}^p$ of s-sparse vectors such that the Hamming distance between any two elements is at least $\frac{s}{2}$ [4,11]. Furthermore, we may compute the pairwise KL divergences in terms of the squared ℓ_2-norm between parameter vectors, so

$$D_{KL}(P_j \| P_k) = \frac{1}{2\sigma^2} \|X(\beta_j - \beta_k)\|_2^2 \leq \frac{4n\delta^2\gamma_{2s}^2}{\sigma^2},$$

where $\gamma_{2s} = \sup_{\beta \in \mathbb{B}_0(2s)} \frac{\|X\beta\|_2}{\sqrt{n}\|\beta\|_2}$. Note that P_j and P_k refer to the conditional distributions of the y_i's given the x_i's for this example, so we are assuming the design matrix is fixed. Applying Theorem 1 and Lemma 2 with ρ equal to the ℓ_2-distance and Φ equal to the identity, we therefore have

$$\inf_{\widehat{\beta}} \sup_{\beta \in \mathbb{B}_0(s) \cap \mathbb{B}_2(1)} \mathbb{E}\left[\|\widehat{\beta} - \beta\|_2\right] \geq \frac{\delta}{2}\left(1 - \frac{\frac{n\delta^2\gamma_{2s}^2}{\sigma^2} - 1}{\frac{s}{2}\log\left(\frac{p-s}{s/2}\right)}\right).$$

Taking $\delta^2 \asymp \frac{\sigma s \log\left(\frac{p-s}{s/2}\right)}{\gamma_{2s}^2 n}$ *and assuming that the problem dimensions satisfy* $n \geq Cs \log p$*, we then obtain a lower bound of the form*

$$\inf_{\widehat{\beta}} \sup_{\beta \in \mathbb{B}_0(s) \cap \mathbb{B}_2(1)} \mathbb{E}\left[\|\widehat{\beta} - \beta\|_2\right] \geq \frac{\delta}{4} \geq \frac{c}{\gamma_{2s}^2} \sqrt{\frac{s}{n} \log\left(\frac{p-s}{s}\right)}.$$

In the case of the ℓ_2*-loss, the Lasso estimator achieves the risk expression in the lower bound (up to constant factors), implying that it is a rate-optimal estimator [4]. Similar bounds on the minimax risk may be derived when the norms appearing in the loss function and/or parameter space are replaced by a general* ℓ_q*-norm [4,12].*

2.3. Metric Entropy

The second method for bounding $I(Y;X)$, due to Yang and Barron [13], is based on the metric entropy of the parameter space. Recall the notion of the *ϵ-covering number* of a set in a metric space, which is the minimum number of ϵ-balls required to cover the set. The logarithm of the covering number is also known as the metric entropy. In particular, we are interested in the quantity $\log N_{KL}(\epsilon; \mathcal{P})$, defined by

$$N_{KL}(\epsilon; \mathcal{P}) = \min\left\{N : \exists\{Q_1, \ldots, Q_N\} \subseteq \mathcal{P} \text{ s.t. } \min_{1 \leq i \leq N} \sqrt{D_{KL}(P, Q_i)} \leq \epsilon, \forall P \in \mathcal{P}\right\},$$

which denotes the ϵ-covering number of $\mathcal{P}$, where distances are measured with respect to the square root KL divergence. We have the following bound:

Lemma 3. *Let* X *and* Y *be defined as in Theorem 1. Then,*

$$I(Y;X) \leq \inf_{\epsilon > 0}\left\{\epsilon^2 + \log N_{KL}(\epsilon; \mathcal{P})\right\}.$$

Proof (sketch). Suppose $\{Q_1, \ldots, Q_N\}$ is an ϵ-cover of $\mathcal{P}$ with respect to the square root KL divergence. Letting $\overline{P} = \frac{1}{M}\sum_{i=1}^{M} P_i$ and $\overline{Q} = \frac{1}{N}\sum_{j=1}^{N} Q_j$, we can check that

$$I(Y;X) = \frac{1}{M}\sum_{i=1}^{M} D_{KL}(P_i \| \overline{P}) \leq \frac{1}{M}\sum_{i=1}^{M} D_{KL}(P_i \| \overline{Q}),$$

where the inequality holds because $\overline{P}$ minimizes the average KL divergence with respect to the second argument. Furthermore, we know that there exists some Q_n such that $D_{KL}(P_i \| Q_n) \leq \epsilon^2$, implying that

$$\begin{aligned} D_{KL}(P_i \| \overline{Q}) &= \int \log \frac{dP_i(X)}{d\overline{Q}(X)} dP_i(X) \leq \int \log \frac{dP_i(X)}{\frac{1}{N} dQ_n(X)} dP_i(X) = D_{KL}(P_i \| Q_n) + \log N \\ &\leq \epsilon^2 + \log N_{KL}(\epsilon; \mathcal{P}). \end{aligned}$$

Since the above inequality holds for all $\epsilon > 0$, we may take an infimum over ϵ to obtain the stated bound. □

As an example of the above technique, we consider the problem of nonparametric regression. Note that the following example shows that the general machinery developed above, though described in terms of parameter estimation, may be applied to nonparametric settings involving function estimation, as well.

Example 2. *(Nonparametric regression) Suppose we observe i.i.d. pairs* $\{(x_i, y_i)\}_{i=1}^n$*, where*

$$y_i = f^*(x_i) + w_i,$$

$x_i \sim Uniform[0,1]$, $w_i \sim N(0,1)$, and x_i is independent of w_i. We also assume that f^* belongs to the function class $\mathcal{F}_s$, for a positive integer s, defined as the set of all continuous functions f on $[0,1]$ satisfying the following properties:

(i) *f is differentiable $s-1$ times on $(0,1)$,*
(ii) *$\sup_{0\le x\le 1} |f^{(k)}(x)| \le 1$, for all $k = 0,1,\ldots,s-1$, where $f^{(0)}(x) := f(x)$,*
(iii) *$f^{(s-1)}$ is 1-Lipschitz on $(0,1)$.*

We derive lower bounds on the minimax risk of estimating f^* when ℓ is the squared L_2-distance, defined by

$$\ell(f,g) = \int_0^1 (f(x) - g(x))^2 \, dx.$$

Hence, we will take $\Phi(t) = t^2$ and ρ equal to the L_2-distance. Let $\mathcal{P}$ denote the set of joint distributions of (x,y) generated by the class $\mathcal{F}_s$. By standard results on the metric entropy of function classes [14,15], we have the bound

$$c\left(\frac{1}{\epsilon}\right)^{1/s} \le \log N_2(\epsilon; \mathcal{F}_s) \le C\left(\frac{1}{\epsilon}\right)^{1/s},$$

where $\log N_2(\epsilon; \mathcal{F}_s)$ denotes the metric entropy of $\mathcal{F}_s$ with respect to the L_2-distance. Furthermore, for any $\delta > 0$, there exists a δ-packing $\{f_1, \ldots, f_M\}$ of $\mathcal{F}_s$ in the L_2-metric such that $\log M = c'\left(\frac{1}{\delta}\right)^{1/s}$. For two functions $f, g \in \mathcal{F}_s$, we may compute the KL divergence between the corresponding distributions $P_f, P_g \in \mathcal{P}$:

$$D_{KL}(P_f \| P_g) = \frac{n}{2} \cdot \|f - g\|_2^2.$$

Hence, it follows that

$$\log N_{KL}(\epsilon; \mathcal{P}) \le \log N_2\left(\epsilon\sqrt{\frac{2}{n}}; \mathcal{F}_s\right) \le C\left(\frac{1}{\epsilon}\sqrt{\frac{n}{2}}\right)^{1/s}.$$

Minimizing the bound obtained from Lemma 3 with respect to ϵ, we obtain $\epsilon^* = C' n^{\frac{1}{4s+2}}$, and plugging back into Theorem 1, we obtain the lower bound

$$\delta^2\left(1 - \frac{C'' n^{1/(4s+2)}}{(1/\delta)^{1/s}}\right).$$

Taking $\delta \asymp \left(\frac{1}{n}\right)^{s/(4s+2)}$ then yields the bound

$$\inf_{\widehat{f}} \sup_{f^* \in \mathcal{F}_s} \mathbb{E}_{f^*}\left[\|\widehat{f} - f^*\|_2^2\right] \ge c'\left(\frac{1}{n}\right)^{s/(2s+1)}.$$

A matching upper bound may be derived using local weighted polynomial regression [16], so the minimax risk is $\Theta\left(n^{-s/(2s+1)}\right)$.

3. Community Recovery

Another area of machine learning that has recently received a substantial amount of attention concerns recovering communities based on node connectivity in a network. A popular probabilistic model is known as the stochastic block model (SBM). In the simplest form of the model, parametrized by (n, K, p, q), the graph has nodes $\{1, \ldots, n\}$ partitioned into K communities. Let the community label of node i be denoted by $\sigma(i)$. The edge set E of the random graph G is then constructed in the following manner: each edge (i,j) is generated independently from all others, with probability

$$P\Big((i,j) \in E\Big) = \begin{cases} p, & \text{if } \sigma(i) = \sigma(j), \\ q, & \text{if } \sigma(i) \neq \sigma(j). \end{cases}$$

The goal is to partition the n nodes into the underlying communities based on observing the graph G.

In order to measure the performance of an algorithm, we consider the loss function

$$r(\widehat{\sigma}, \sigma) = \frac{1}{n} \min_{\tau \in S_K} d_H(\widehat{\sigma}, \tau \circ \sigma).$$

Here, the estimator $\widehat{\sigma} : \{1, \ldots, n\} \to \{1, \ldots, K\}$ corresponds to a partitioning of the nodes into K communities, and d_H denotes the Hamming distance between assignments. Furthermore, we take the minimum over all permutations S_K of the community labels. Hence, $r(\widehat{\sigma}, \sigma)$ is the proportion of incorrectly labeled nodes (for the optimal labeling of partitions). We will focus our discussion on the setting where K is fixed, but p and q may vary with n; generalizations exist in the literature where K is allowed to grow with n, as well. We are interested in the behavior of various algorithms as $n \to \infty$.

In the following two subsections, we discuss the popular notions of *weak recovery* and *exact recovery*. The algorithm $\widehat{\sigma}$ achieves weak recovery if $\mathbb{E}[r(\widehat{\sigma}, \sigma)] \to 0$ (i.e., the expected fraction of misclassified nodes tends to 0 as $n \to \infty$), and achieves exact recovery if $r(\widehat{\sigma}, \sigma) = 0$. For a more complete description of current work on stochastic block models, see the extensive survey paper by Abbe [17].

3.1. Weak Recovery

Analogous to the setting discussed in Section 2, we may derive bounds on the minimax risk

$$\inf_{\widehat{\sigma}} \sup_{\sigma \in \Sigma(n,K)} \mathbb{E}[r(\widehat{\sigma}, \sigma)],$$

where $\Sigma(n, K)$ is an appropriate class of underlying community labelings. We state and prove a result for approximately equal-sized communities in the limit as $n \to \infty$, so $\Sigma(n, K)$ is the set of all labelings σ such that $|\{i : \sigma(i) = k\}| = (1 + o(1))\frac{n}{K}$, for all $1 \leq k \leq K$.

The main result is the following [18]:

Theorem 2. *Suppose $p = \frac{a}{n}$ and $q = \frac{b}{n}$, and suppose $\frac{nI}{K} \to \infty$, where*

$$I = -2 \log \left(\sqrt{\frac{a}{n}} \sqrt{\frac{b}{n}} + \sqrt{1 - \frac{a}{n}} \sqrt{1 - \frac{b}{n}} \right). \tag{5}$$

A lower bound on the minimax risk of community estimation is given by

$$\inf_{\widehat{\sigma}} \sup_{\sigma \in \Sigma(n,K)} \mathbb{E}[r(\widehat{\sigma}, \sigma)] \geq \exp\left(-(1 + o(1)) \frac{nI}{K} \right).$$

Proof (sketch). The core of the approach bears similarity to the method for obtaining lower bounds for estimation, in the sense that we construct a subset Σ^L of the parameter space corresponding to "messages", which we wish to recover via an appropriate decoding strategy. In the case when $K = 2$ (and n is even), the subset Σ^L consists of all partitions of the nodes into equal-sized communities and communities of size $\left(\frac{n}{2} + 1, \frac{n}{2} - 1\right)$. We focus on the case $K = 2$ in the present proof sketch to avoid technical complications.

The proof is somewhat more involved than the strategies outlined in Section 2, however, since the unknown quantity to be estimated is a set of discrete labelings and the loss function is defined with respect to an optimal permutation. The first step is to lower-bound the minimax risk by the average risk over the class Σ^L. Furthermore, a more technical argument shows that we may just examine the average local risk defined with respect to a single node in the graph:

$$\begin{aligned}\inf_{\widehat{\sigma}} \sup_{\sigma \in \Sigma(n,K)} \mathbb{E}[r(\widehat{\sigma}, \sigma)] &\geq \inf_{\widehat{\sigma}} \sup_{\sigma \in \Sigma^L} \mathbb{E}_{\sigma}[r(\widehat{\sigma}, \sigma)] \\ &\geq \inf_{\widehat{\sigma}} \frac{1}{|\Sigma^L|} \sum_{\sigma \in \Sigma^L} \mathbb{E}[r(\widehat{\sigma}, \sigma)] \\ &= \inf_{\widehat{\sigma}} \frac{1}{|\Sigma^L|} \sum_{\sigma \in \Sigma^L} \mathbb{E}[r_1(\widehat{\sigma}, \sigma)],\end{aligned}$$

where r_1 is the local loss function defined with respect to node 1, which is the fraction of optimal permutations of community assignments that incorrectly classify node 1. The next step is to lower-bound the local risk (uniformly over all choices of $\sigma \in \Sigma^L$) using the minimum risk of a binary hypothesis testing problem, where the two hypotheses correspond to the possible assignments of node 1 as a member of the first or second community. In particular, we have the following inequality, which holds for each σ:

$$\mathbb{E}[r_1(\widehat{\sigma}, \sigma)] \geq c\, \mathbb{P}\left(\sum_{i=1}^{n/2} X_i \geq \sum_{j=1}^{n/2} Y_j \right),$$

where $X_i \overset{i.i.d.}{\sim} \textit{Bernoulli}\left(\frac{b}{n}\right)$ and $Y_j \overset{i.i.d.}{\sim} \textit{Bernoulli}\left(\frac{a}{n}\right)$ are independent random variables. Standard techniques involving large deviation inequalities allow us to lower-bound the latter probability, thus yielding the overall lower bound appearing in the theorem. □

As demonstrated by Zhang and Zhou [18], the lower bound on the risk appearing in Theorem 2 may be achieved using a form of penalized likelihood estimation. A computationally feasible procedure was subsequently provided in Gao et al. [19].

Remark 1. *The quantity I appearing in Equation (5) is the Renyi divergence of order $\frac{1}{2}$ between a Bernoulli$\left(\frac{a}{n}\right)$ and Bernoulli$\left(\frac{b}{n}\right)$ distribution. In fact, these results generalize to the case of non-binary edge weights, and the Renyi divergence of order $\frac{1}{2}$ also appears in the minimax rates for estimation in weighted stochastic block models [20]. Furthermore, if the communities are not all of equal size, alternative divergence functions appear in the error exponent [21,22]. Finally, note that the regime where $p = \frac{a}{n}$ and $q = \frac{b}{n}$, with $a, b = \Theta(1)$, corresponds to the threshold at which giant components emerge in the network [23]. Theorem 2 allows a and b to scale arbitrarily with n, provided $\frac{nI}{K} \to \infty$, which will not hold if $a, b \ll n$.*

3.2. Exact Recovery

Information-theoretic arguments may also be used to establish lower bounds for exact recovery in stochastic block models, which corresponds to correct classification of every single node (up to permutation the of community labels). We present a result, due to Abbe et al. [24], that provides lower bounds for exact recovery in the case of two equal-sized communities.

We have the following result:

Theorem 3. *Let $p = \frac{a \log n}{n}$ and $b = \frac{q \log n}{n}$, where $a > b \geq 0$. If $\left(\sqrt{a} - \sqrt{b}\right)^2 < 2$, then for sufficiently large n, the maximum likelihood estimator fails in recovering the communities with probability bounded away from 0:*

$$\liminf_{n \to \infty} \mathbb{P}\left(r(\widehat{\sigma}_{MLE}, \sigma) \neq 0\right) > 0.$$

Proof (sketch). We denote the two communities by A and B. Let F be the event that the maximum likelihood estimator fails in performing exact recovery, and let

$$\begin{aligned}F_A &= \{\exists i \in A : i \text{ is connected to more nodes in } B \text{ than in } A\}, \\ F_B &= \{\exists j \in B : j \text{ is connected to more nodes in } A \text{ than in } B\}.\end{aligned}$$

By symmetry, we have $\mathbb{P}(F_A) = \mathbb{P}(F_B)$. Furthermore, note that

$$F_A \cap F_B \subseteq F,$$

since if both F_A and F_B were to occur simultaneously, swapping the labels of the nodes i and j would lead to a higher value of the likelihood than in the case of correct labeling. In particular, this implies that

$$\mathbb{P}(F) \geq \mathbb{P}(F_A \cap F_B) \geq \mathbb{P}(F_A) + \mathbb{P}(F_B) - 1 = 2\mathbb{P}(F_A) - 1. \tag{6}$$

Let $H \subseteq A$ denote a fixed subset with $|H| = \left\lfloor \frac{n}{\log^3(n)} \right\rfloor$, and define the event

$$F_H = \left\{ \exists j \in H \text{ s.t. } E(j, A \backslash H) + \frac{\log n}{\log\log n} \leq E(j, B) \right\},$$

where $E(j, C)$ denotes the number of edges between j and the nodes in C. Note that, if event F_H occurs and all nodes in H are connected to at most $\frac{\log n}{\log\log n}$ other nodes in H, then event F_A must occur. Furthermore, one can show that, with high probability, every node in H is connected to at most $\frac{\log n}{\log\log n}$ other nodes in H. Hence,

$$\mathbb{P}(F_A) \geq \mathbb{P}(F_H) + o(1). \tag{7}$$

It remains to derive a lower bound on $\mathbb{P}(F_H)$. For $j \in H$, let

$$F_H^{(j)} = \left\{ E(j, A \backslash H) + \frac{\log n}{\log\log n} \leq E(j, B) \right\},$$

and note that the $F_H^{(j)}$'s are independent. Hence,

$$\mathbb{P}(F_H) = \mathbb{P}\left(\bigcup_{j \in H} F_H^{(j)} \right) = 1 - \prod_{j \in H} \left(1 - \mathbb{P}(F_H^{(j)}) \right).$$

Straightforward techniques for bounding sums of independent Bernoulli random variables show that $\mathbb{P}(F_H^{(j)}) > \frac{\log(4)\log^3(n)}{n}$ for each j, from which we can conclude that

$$\mathbb{P}(F_H) \geq 1 - \left(1 - \frac{\log(4)\log^3(n)}{n} \right)^{\left\lfloor \frac{n}{\log^3(n)} \right\rfloor} = 1 - \frac{1}{4} + o(1). \tag{8}$$

Combining inequalities (6)–(8) then yields the desired result. □

Note that for any other estimator $\widehat{\sigma}$, we have

$$\mathbb{P}\left(r(\widehat{\sigma}_{MLE}, \sigma) = 0 \right) \geq \mathbb{P}\left(r(\widehat{\sigma}, \sigma) = 0 \right).$$

Hence, Theorem 3 also implies that

$$\liminf_{n \to \infty} \mathbb{P}\left(r(\widehat{\sigma}, \sigma) \neq 0 \right) \geq \mathbb{P}\left(r(\widehat{\sigma}_{MLE}, \sigma) \neq 0 \right) \geq 0.$$

In fact, a converse of Theorem 3 holds, as well:

Theorem 4. *Under the same conditions as in Theorem 3, suppose instead that* $\left(\sqrt{a}-\sqrt{b}\right)^2 > 2$. *Then, the maximum likelihood estimator succeeds in recovering the communities with probability tending to 1:*

$$\lim_{n\to\infty} \mathbb{P}\left(r(\widehat{\sigma}_{MLE}, \sigma) = 0\right) = 1.$$

Since the focus of this paper is to establish lower bounds, we refer the reader to Abbe [24] for the proof of Theorem 4, which proceeds by direct calculation. An extension of Theorems 3 and 4 for weighted stochastic block models may be found in Jog and Loh [25].

Remark 2. *The threshold behavior described in Theorems 3 and 4 is perhaps not surprising in light of known threshold behavior in Shannon coding theory, and the connections between each of the statistical learning tasks and the problem of decoding on a discrete alphabet after passage through a noisy channel. Indeed, the community recovery problem has been cast in information-theoretic terminology as decoding in a "graphical channel" [26]. On the other hand, the coding scheme is fixed according to the stochastic block model, whereas Shannon theory allows one to design an optimal encoding scheme to achieve channel capacity. See also the paper by Chen et al. [27], and the derivation of similar types of sharp threshold behavior in submatrix localization problems [28,29]. Finally, we note that the scaling* $p = \frac{a\log n}{n}$ *and* $q = \frac{b\log n}{n}$, *when* $a, b = \Theta(1)$, *corresponds to the threshold for the graph to have isolated vertices with probability tending to 1 [23]. Indeed, it would be impossible to perform exact recovery with high probability in the presence of isolated vertices: flipping the community assignments of two isolated vertices belonging to the two different communities would not change the value of the likelihood.*

4. Online Learning

We now shift our focus to sequential allocation problems. The setup we consider involves a series of actions taken by a player, using limited feedback about the environment based on his/her past actions. We study the setting of a *multi-armed bandit*, where each potential action of the player is associated with a reward distribution, but the player only observes the reward corresponding to his/her action on successive rounds. In the following two subsections, we will consider the cases of *stochastic* and *adversarial* bandits and obtain bounds on a quantity known as regret. More details on the setting and results may be found in Bubeck and Cesa-Bianchi [30] or Cesa-Bianchi and Lugosi [31].

4.1. Stochastic Bandits

We first analyze the setting of stochastic multi-armed bandits. On each round, the player may choose one of k different arms. Associated to arm j is a reward distribution P_{θ_j}, where $\theta_j \in \Theta$ belongs to some parameter space. Furthermore, we assume that the reward distributions $\{P_{\theta_j}\}_{j=1}^k$ remain fixed across all rounds. We use the notation $\mu(\theta)$ to denote the mean of the distribution P_θ, and let $\mu^* = \max_{1\le j\le k} \mu(\theta_j)$ denote the maximum expected reward.

Denote the sequence of actions chosen by the player as $(I_1, \ldots, I_n)$, where $I_t \in \{1, \ldots, k\}$ is the arm played at time t, and let $X_{I_t,t} \sim P_{\theta_{I_t}}$ denote the observed reward, which is an i.i.d. drawn from the distribution $P_{\theta_{I_t}}$. Note that I_t may be a function of the previously observed reward sequence $(X_{I_1,1}, \ldots, X_{I_{t-1},t})$ and may also involve additional randomization. We are interested in bounding a quantity known as the *pseudo-regret*, defined as

$$\overline{R}_n = n\mu^* - \mathbb{E}\left[\sum_{t=1}^n X_{I_t,t}\right],$$

where we may also write $\overline{R}_n(\theta_1, \ldots, \theta_k)$ to make the dependence on the reward distributions explicit. If the player employs a random strategy, the expectation is computed with respect to randomness in the sequence of actions $(I_1, \ldots, I_n)$, as well as randomness generated by draws from the reward distributions. In other words, the pseudo-regret measures the difference between the expected reward

incurred by the player's strategy and the expected reward incurred by playing the arm with maximum expected reward on every round.

Lai and Robbins [32] prove the following result. We omit some technical regularity conditions on the parameter space, such as denseness of the parameter space and continuity with respect to the KL divergence, in order to avoid cluttering the presentation.

Theorem 5. *Suppose that, for all pairs $\theta_1, \theta_2 \in \Theta$ such that $\mu(\theta_1) > \mu(\theta_2)$, we have $0 < D_{KL}(P_{\theta_2} \| P_{\theta_1}) < \infty$. Suppose a strategy satisfies $\overline{R}_n(\theta_1, \ldots, \theta_k) = o(n^\alpha)$, for all $\theta_1, \ldots, \theta_k \in \Theta$ and all $\alpha > 0$. Then, for any $(\theta_1, \ldots, \theta_k) \in \Theta$, we have*

$$\liminf_{n\to\infty} \frac{\overline{R}_n(\theta_1, \ldots, \theta_k)}{\log n} \geq \sum_{j:\mu_j<\mu^*} \frac{\mu^* - \mu_j}{D_{KL}(P_{\theta_j} \| P_{\theta^*})},$$

where $\theta^* \in \arg\min_{\theta\in\Theta} \mu(\theta)$ and $\mu_j = \mu(\theta_j)$.

Proof (sketch). We may write

$$\overline{R}_n = \sum_{j=1}^{k} \mathbb{E}[T_j(n)]\Delta_j,$$

where $\Delta_j = \mu^* - \mu_j$ and $T_j(n) = \sum_{t=1}^{n} 1\{I_t = j\}$. The main step is to show that the inequality

$$\mathbb{E}[T_j(n)] \geq \frac{\log n}{D_{KL}(P_{\theta_j} \| P_{\theta^*})}, \qquad \forall j : \mu_j < \mu^* \tag{9}$$

holds for any strategy. Inequality (9) provides a lower bound on the expected number of pulls to any suboptimal arm (note that, as P_{θ_j} becomes further from P_{θ^*}, the two arms are easier to distinguish, so the expected number of pulls to the suboptimal arm can be smaller). We focus on proving inequality (9) for $j = 2$; the other cases are similar.

Consider two parameter vectors $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ and $\theta' = (\theta_1, \theta_2', \ldots, \theta_k)$, which differ only in the second coordinate. We further choose the parameters such that

$$\mu_1 > \mu_2 \geq \mu_3 \geq \cdots \geq \mu_k,$$
$$\mu_2' \geq \mu_1 > \mu_3 \geq \cdots \geq \mu_k,$$

so the second arm is suboptimal in the first setting but optimal in the second. We will choose θ_2' close to θ_1, so

$$D_{KL}(P_{\theta_2} \| P_{\theta_2'}) \approx D_{KL}(P_{\theta_2} \| P_{\theta_1}) = D_{KL}(P_{\theta_2} \| P_{\theta^*}).$$

(The regularity conditions on the parameter space and reward distributions ensure that such a choice is possible.) The idea is that, since P_θ and $P_{\theta'}$ are close, any strategy should pick roughly the same sequence of arms in both scenarios, but a strategy that performs well on θ will behave relatively poorly on θ' (and vice versa), since the ordering of arms according to optimality is different in the two settings. In particular, we will derive the following bound, relating the probabilities of pulling the second arm in each of the parameter settings:

$$\mathbb{P}_\theta(T_2(n) < a_n) \leq c_n' \, \mathbb{P}_{\theta'}(T_2(n) < a_n) + b_n, \tag{10}$$

where $a_n = \frac{(1-3\alpha)\log n}{D_{KL}(P_{\theta_2} \| P_{\theta_2'})}$, and we take $\alpha < \frac{1}{3}$. We can show that $b_n = o(1)$ since P_θ and $P_{\theta'}$ are close, and that the right-hand probability is also $o(1)$, since arm 2 is optimal under θ'.

For a fixed strategy, let $\{X_{j,s}\}_{\substack{1\leq j\leq k \\ 1\leq s\leq n}}$ denote the rewards corresponding to various arm pulls. For $A \subseteq \{T_2(n) = n_2\}$, we have

$$\begin{aligned}\mathbb{P}_{\theta'}(A) &= \int_A d\mathbb{P}_{\theta'}(x) = \int_A \frac{d\mathbb{P}_{\theta'}(x)}{d\mathbb{P}_{\theta}(x)} \cdot d\mathbb{P}_{\theta}(x) \\ &= \int_A \prod_{s=1}^{n_2} \frac{dP_{\theta_2'}(x_{2,s})}{dP_{\theta_2}(x_{2,s})} d\mathbb{P}_{\theta}(x) \\ &= \int_A e^{-L_2(x)} d\mathbb{P}_{\theta}(x),\end{aligned}$$

where we define $L_2(x) = \sum_{s=1}^{T_2(n)} \log \frac{dP_{\theta_2}(x_{2,s})}{dP_{\theta_2'}(x_{2,s})}$. In particular, if $A \subseteq \{T_2(n) = n_2, L_2(X) \leq c_n\}$, we have

$$\mathbb{P}_{\theta'}(A) \geq e^{-c_n}\, \mathbb{P}_{\theta}(A),$$

where we will take $c_n = (1-2\alpha)\log n$. We may therefore write

$$\begin{aligned}\mathbb{P}_{\theta}(T_2(n) < a_n) &= \mathbb{P}_{\theta}(T_2(n) < a_n, L_2(X) \leq c_n) + \mathbb{P}_{\theta}(T_2(n) < a_n, L_2(X) > c_n) \\ &\leq e^{c_n}\, \mathbb{P}_{\theta'}(T_2(n) < a_n, L_2(X) \leq c_n) + \mathbb{P}_{\theta}(T_2(n) < a_n, L_2(X) > c_n) \\ &\leq e^{c_n}\, \mathbb{P}_{\theta'}(T_2(n) < a_n) + \mathbb{P}_{\theta}(T_2(n) < a_n, L_2(X) > c_n),\end{aligned}$$

which is inequality (10) with $c_n' = e^{c_n}$ and $b_n = \mathbb{P}_{\theta}(T_2(n) < a_n, L_2(X) > c_n)$. Note that if $T_2(n) < a_n$, we have $L_2(X) < \sum_{s=1}^{a_n} \log \frac{dP_{\theta_2}(X_{2,s})}{dP_{\theta_2'}(X_{2,s})}$, so

$$b_n \leq \mathbb{P}_{\theta}\left(\sum_{s=1}^{a_n} \log \frac{dP_{\theta_2}(X_{2,s})}{dP_{\theta_2'}(X_{2,s})} > c_n\right) = o(1),$$

where the last equality follows from the fact that the rewards $\{X_{2,s}\}_{s=1}^{a_n}$ are i.i.d. and

$$\frac{1}{a_n}\sum_{s=1}^{a_n} \log \frac{dP_{\theta_2}(X_{2,s})}{dP_{\theta_2'}(X_{2,s})} \xrightarrow{a.s.} \mathbb{E}_{\theta}\left[\log \frac{dP_{\theta_2}(X_{2,s})}{dP_{\theta_2'}(X_{2,s})}\right] = D_{KL}(P_{\theta_2} \| P_{\theta_2'}).$$

Finally, we bound $\mathbb{P}_{\theta'}(T_2(n) < a_n)$ using Markov's inequality:

$$\mathbb{P}_{\theta'}(T_2(n) < a_n) = \mathbb{P}_{\theta'}(n - T_2(n) \geq n - a_n) \leq \frac{\mathbb{E}_{\theta'}[n - T_2(n)]}{n - a_n} = o(n^{\alpha-1}),$$

where the last equality follows from the fact that $a_n = o(n)$ and the assumption on $\overline{R}_n(\theta')$. Altogether, we conclude that the right-hand side of inequality (10) is $o(1)$.

By another application of Markov's inequality, we conclude that

$$\mathbb{E}_{\theta}[T_2(n)] \cdot \frac{D_{KL}(P_{\theta_2} \| P_{\theta_2'})}{\log n} \geq \mathbb{P}_{\theta}\left(T_2(n) \geq \frac{\log n}{D_{KL}(P_{\theta_2} \| P_{\theta_2'})}\right) > \mathbb{P}_{\theta}(T_2(n) > a_n) \to 1.$$

Hence,

$$\frac{\mathbb{E}_{\theta}[T_2(n)]}{\log n} \geq \frac{1}{D_{KL}(P_{\theta_2} \| P_{\theta_2'})} \approx \frac{1}{D_{KL}(P_{\theta_2} \| P_{\theta^*})},$$

as wanted. □

Note that the assumption $\overline{R}_n(\theta_1, \ldots, \theta_k) = o(n^{\alpha})$ implies that a sufficiently good player strategy exists for all choices of reward parameters. In particular, such a condition may be verified when the reward distributions are Bernoulli (e.g., $P_{\theta} \sim \text{Bernoulli}(\theta)$). Then, we have

$$D_{KL}(P_{\theta_1} \| P_{\theta_2}) = \theta_1 \log \frac{\theta_1}{\theta_2} + (1-\theta_1) \log \frac{1-\theta_1}{1-\theta_2},$$

and combined with Theorem 5, we obtain the lower bound

$$\liminf_{n\to\infty} \frac{\overline{R}_n(\theta_1,\ldots,\theta_k)}{\log n} \geq \mu^*(1-\mu^*) \sum_{j:\mu_j<\mu^*} \frac{1}{\mu^*-\mu_j}.$$

A player strategy known as the Upper Confidence Bound (UCB) strategy may be shown to achieve this lower bound, up to constant factors [32,33].

Finally, we mention a non-asymptotic lower bound on the pseudo-regret that comes from the probably approximately correct (PAC) literature on bandits [34–36]:

Theorem 6. *In the case of Bernoulli reward distributions, there exist positive constants* $\{c_i\}_{i=1}^5$ *such that for all* $k \geq 2$ *and* $n \geq 1$*, the pseudo-regret of any strategy satisfies*

$$\sup_{\theta_1,\ldots,\theta_k\in[0,1]} \overline{R}_n(\theta_1,\ldots,\theta_k) \geq \min\left\{c_1 n,\ c_2 k + c_3 n,\ c_4 k(\log n - \log k + c_5)\right\}. \tag{11}$$

Proof (sketch). For a detailed proof of Theorem 6, we refer the reader to Mannor and Tsitsiklis [36]. The main idea is to construct a collection of k vectors $\{\theta^1,\ldots,\theta^k\} \subseteq [0,1]^k$ corresponding to the parameters of the reward distributions on arms. For each $2 \leq i \leq k$, we define the vector $\theta^i = (\theta^i_1,\ldots,\theta^i_k)$ such that

$$\theta^i_1 = \frac{1}{2} + \frac{\epsilon}{2}, \qquad \theta^i_i = \frac{1}{2} + \epsilon, \qquad \theta^i_j = \frac{1}{2}, \text{ for } j \notin \{1,i\},$$

and we define the vector θ^1 such that

$$\theta^1_1 = \frac{1}{2} + \frac{\epsilon}{2}, \qquad \theta^1_j = \frac{1}{2}, \text{ for } j > 1.$$

In other words, the reward distribution of arm 1 is the same for all k parameter settings, but in the case of vector θ^i, the reward distribution for arm i is slightly better than the reward distributions of the other arms. We then compute a weighted sum of the regret incurred in each parameter setting, where θ^1 is given weight $\frac{1}{2}$ and all other θ^i's are given weight $\frac{1}{2(n-1)}$. We may show that this weighted regret is lower-bounded by the quantity appearing in inequality (11), implying the existence of at least one parameter setting that satisfies the desired bound. Computing the lower bound for the weighted regret is similar to the procedure adopted in the proof of Theorem 5, in that we compute a lower bound on the expected number of arm pulls of each suboptimal arm in each parameter setting in terms of ϵ. □

Theorem 6 is a type of minimax result, stating that, for *any* player strategy, a distribution of Bernoulli rewards exists for which the problem incurs $\Omega(\log n)$ regret. The same UCB strategies of Auer et al. [33] may be used to obtain $\mathcal{O}(\log n)$ upper bounds on the minimax regret even for the worst-case reward distribution, showing that the bound stated in Theorem 6 is tight.

4.2. Adversarial Bandits

In the adversarial setting, we allow the reward distributions to vary arbitrarily over time. Thus, we assume that the reward distributions are chosen by an "adversary", where the class of permissible adversarial strategies is denoted by $\mathcal{P}$. For a player strategy S and an adversarial strategy $P \in \mathcal{P}$, we define the pseudo-regret analogously to the stochastic case:

$$\overline{R}_n(S,P) = \max_{1\leq j\leq k} \mathbb{E}_P\left[\sum_{t=1}^n X_{j,t}\right] - \mathbb{E}_{S,P}\left[\sum_{t=1}^n X_{I_t,t}\right],$$

where the first expectation is taken with respect to possible randomization in the adversarial strategy, and the second expectation is taken with respect to randomization in the strategies of both the player and adversary.

The following result provides a lower bound for the minimax pseudo-regret, where the supremum is taken over $\mathcal{P}_{Ber}$, the set of all Bernoulli reward distributions over the k time steps, and the infimum is taken over all player strategies [30,37]:

Theorem 7. *The minimax pseudo-regret satisfies the bound*

$$\inf_{S \in \mathcal{S}} \sup_{P \in \mathcal{P}_{Ber}} \overline{R}_n(S, P) \geq \frac{1}{18} \cdot \min\{\sqrt{nk}, n\},$$

where the infimum is taken over all (possibly randomized) player strategies.

Proof (sketch). Note that it suffices to prove the bound when the infimum is taken over deterministic player strategies, since the pseudo-regret for a randomized strategy will be a convex combination of the pseudo-regret of deterministic strategies. Fix a deterministic player strategy, and consider the reward distributions $\mathbb{P}_1, \ldots, \mathbb{P}_k \in \mathcal{P}_{Ber}$, where $\mathbb{P}_j$ corresponds to the distribution where the reward of each arm $i \neq j$ is i.i.d. Bernoulli$\left(\frac{1}{2}\right)$, and the reward of arm j is i.i.d. Bernoulli$\left(\frac{1}{2} + \epsilon\right)$. Note that this construction bears some similarity to the proof outline for Theorem 6 provided above, in that the reward distribution $\mathbb{P}_j$ slightly favors arm j. We will also compute a lower bound for the weighted regret, this time allocating uniform weights to each parameter setting, in order to conclude the existence of at least one assignment of reward distributions satisfying the desired lower bounds. Let $\mathbb{E}_j$ denote the expectation with respect to the reward distribution $\mathbb{P}_j$.

We may compute

$$\frac{1}{k} \sum_{j=1}^{k} \overline{R}_n(S, \mathbb{P}_j) = \frac{1}{k} \sum_{j=1}^{k} \mathbb{E}_j \left[\sum_{i \neq j} \epsilon T_i(n) \right] = \frac{\epsilon}{k} \sum_{j=1}^{k} \mathbb{E}_j \left[n - T_j(n) \right] = \epsilon \left(n - \frac{1}{k} \sum_{j=1}^{k} \mathbb{E}_j[T_j(n)] \right), \tag{12}$$

where $T_i(n)$ denotes the number of pulls of arm i.

Let $\mathbb{P}$ denote the reward distribution where all arms have a Bernoulli$\left(\frac{1}{2}\right)$ distribution. We may obtain the following bound:

$$\mathbb{E}_j[T_j(n)] \overset{(a)}{\leq} \mathbb{E}_{\mathbb{P}}[T_j(n)] + n\sqrt{\frac{1}{2} D_{KL}(\mathbb{P} \| \mathbb{P}_j)} \overset{(b)}{=} \mathbb{E}_{\mathbb{P}}[T_j(n)] + \frac{n}{2}\sqrt{\log\left(\frac{1}{1-4\epsilon^2}\right) \mathbb{E}[T_j(n)]}, \tag{13}$$

where inequality (a) may be derived by first relating the difference in expectations for bounded random variables to total variation distance and then applying Pinsker's inequality, and equality (b) follows from a direct computation. Combining inequalities (12) and (13), we then obtain

$$\begin{aligned}
\frac{1}{k} \sum_{j=1}^{k} \overline{R}_n(S, \mathbb{P}_j) &\geq \epsilon \left(n - \frac{n}{k} - \frac{n}{2k} \sqrt{\log\left(\frac{1}{1-4\epsilon^2}\right)} \sum_{j=1}^{k} \sqrt{\mathbb{E}[T_j(n)]} \right) \\
&= \epsilon n \left(1 - \frac{1}{k} - \frac{1}{2} \sqrt{\log\left(\frac{1}{1-4\epsilon^2}\right)} \frac{1}{k} \sum_{j=1}^{k} \sqrt{\mathbb{E}[T_j(n)]} \right) \\
&\geq \epsilon n \left(1 - \frac{1}{k} - \frac{1}{2} \sqrt{\log\left(\frac{1}{1-4\epsilon^2}\right)} \sqrt{\frac{n}{k}} \right),
\end{aligned}$$

using the concavity of the square root function. Choosing $\epsilon = \frac{1}{4n}\min\{\sqrt{kn}, n\}$ then yields the inequality

$$\sup_{P \in \mathcal{P}_{Ber}} \overline{R}_n(S, P) \geq \frac{1}{k}\sum_{j=1}^{k} \overline{R}_n(S, \mathbb{P}_j) \geq \frac{1}{18}\min\{\sqrt{kn}, n\},$$

and taking an infimum over all player strategies produces the desired result. □

Note that the lower bound provided in Theorem 7 clearly also holds when the supremum is taken over any class of adversarial strategies containing $\mathcal{P}_{Ber}$. In particular, one topic of study is that of *oblivious* adversaries, which are allowed to perform any strategy that is non-adaptive to the actions of the player (i.e., it is chosen before the start of the first round). The Exp3 algorithm provides an upper bound on the minimax pseudo-regret for oblivious adversaries that matches the lower bound in Theorem 7 up to a factor of $\sqrt{\log k}$ [37]. The study of non-oblivious adversaries refers to the setting where the adversary's actions may be chosen in response to the player's sequential choices, as well, and is also an active area of research [31,38].

5. Discussion

In this article, we have presented several distinct approaches for deriving lower bounds in various statistical learning problems. In each of the settings described—statistical estimation, community recovery, and online learning—we have shown how to simplify the problem to one involving channel decoding, and leverage information-theoretic bounds on the hardness of the decoding problem to bound the hardness of the corresponding statistical problem. It is worth reflecting on the similarities between the techniques employed in each of the approaches. Although the specific interpretation involving channel decoding looks quite different in each of the settings, the trick is to find an appropriate discretization of parameter space so that pairs of parameters are relatively far apart, but the corresponding data-generating distributions are close. In the context of statistical estimation, this means that we construct a packing of parameter space. In the community recovery setting, we consider pairs of community partitions that differ only in the assignment of a single node. In the multi-armed bandit setting, we consider pairs of arm parameters that flip the assignment of the optimal arm, while perturbing the parameter values as little as possible.

On a more applied note, information-theoretic tools have made an appearance in various machine learning algorithms involving maximizing independence between observed quantities. Some examples include decision tree learning via information gain [39]; independent component analysis by mutual information minimization [40]; causal inference algorithms maximizing independence [41]; minimal-redundancy-maximal-relevance (mRMR) methods for feature selection [42]; and image registration via mutual information maximization in medical imaging [43]. As a result, quantities such as mutual information have become increasingly mainstream in data science applications. Note, however, that such applications of information theory to machine learning have no connection to the channel decoding techniques or hardness results discussed in this article. In terms of statistical theory, these applications have created a renewed interest in deriving efficient estimators of entropy and other related information measures based on finite samples [44–47], but a detailed discussion of such methods is somewhat orthogonal to the main topic of this survey.

Acknowledgments: The author thanks Varun Jog, the Assitant Editor, and the anonymous referees for helpful comments that enhanced the clarity of the paper.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Bousquet, O.; Boucheron, S.; Lugosi, G. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*; Springer: Berlin/Heidelberger, Germany, 2004; pp. 169–207.

2. Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: Berlin/Heidelberger, Germany, 2001; Volume 1.
3. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: New York, NY, USA, 2012.
4. Raskutti, G.; Wainwright, M.J.; Yu, B. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inf. Theory* **2011**, *57*, 6976–6994.
5. Tsybakov, A.B. *Introduction to Nonparametric Estimation*; Springer: Berlin/Heidelberger, Germany, 2008.
6. Santhanam, N.P.; Wainwright, M.J. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Trans. Inf. Theory* **2012**, *58*, 4117–4134.
7. Guntuboyina, A. Lower bounds for the minimax risk using f-divergences, and applications. *IEEE Trans. Inf. Theory* **2011**, *57*, 2386–2399.
8. Cai, T.T.; Zhang, C.H.; Zhou, H.H. Optimal rates of convergence for covariance matrix estimation. *Ann. Stat.* **2010**, *38*, 2118–2144.
9. Amini, A.A.; Wainwright, M.J. High-Dimensional Analysis of Semidefinite Relaxations for Sparse Principal Components. *Ann. Stat.* **2009**, *37*, 2877–2921.
10. Lehmann, E.L.; Casella, G. *Theory of Point Estimation*; Springer Science & Business Media: Berlin/Heidelberger, Germany, 2006.
11. Kühn, T. A lower estimate for entropy numbers. *J. Approx. Theory* **2001**, *110*, 120–124.
12. Zhang, C.H. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **2010**, *38*, 894–942.
13. Yang, Y.; Barron, A. Information-theoretic determination of minimax rates of convergence. *Ann. Stat.* **1999**, *27*, 1564–1599.
14. Lorentz, G.G. Metric entropy and approximation. *Bull. Am. Math. Soc.* **1966**, *72*, 903–937.
15. Tikhomirov, V.M.; Shiryayev, A.N. ϵ-entropy and ϵ-capacity of sets in functional spaces. In *Selected Works of A.N. Kolmogorov: Volume III: Information Theory and the Theory of Algorithms*; Springer: Dordrecht, The Netherlands, 1993; pp. 86–170.
16. Stone, C.J. Optimal global rates of convergence for nonparametric regression. *Ann. Stat.* **1982**, *10*, 1040–1053.
17. Abbe, E. Community detection and stochastic block models: Recent developments. *arXiv* **2017**, arXiv:1703.10146.
18. Zhang, A.Y.; Zhou, H.H. Minimax rates of community detection in stochastic block models. *Ann. Stat.* **2016**, *44*, 2252–2280.
19. Gao, C.; Ma, Z.; Zhang, A.Y.; Zhou, H.H. Achieving optimal misclassification proportion in stochastic block model. *arXiv* **2015**, arXiv:1505.03772.
20. Xu, M.; Jog, V.; Loh, P. Optimal Rates for Community Estimation in the Weighted Stochastic Block Model. *arXiv* **2017**, arXiv:1706.01175.
21. Abbe, E.; Sandon, C. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS), Berkeley, CA, USA, 17–20 October 2015; pp. 670–688.
22. Yun, S.Y.; Proutiere, A. Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems, Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, 4–9 December 2016*; The Neural Information Processing Systems (NIPS) Foundation: La Jolla, CA, USA, 2016; pp. 965–973.
23. Bollobás, B. *Random Graphs (Cambridge Studies in Advanced Mathematics)*; Cambridge University Press: Cambridge, UK, 2001.
24. Abbe, E.; Bandeira, A.S.; Hall, G. Exact recovery in the stochastic block model. *IEEE Trans. Inf. Theory* **2016**, *62*, 471–487.
25. Jog, V.; Loh, P. Recovering communities in weighted stochastic block models. In Proceedings of the 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 29 September–2 October 2015; pp. 1308–1315.
26. Abbe, E.; Montanari, A. Conditional random fields, planted constraint satisfaction and entropy concentration. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*; Springer: Berlin/Heidelberger, Germany, 2013; pp. 332–346.
27. Chen, Y.; Suh, C.; Goldsmith, A.J. Information recovery from pairwise measurements. *IEEE Trans. Inf. Theory* **2016**, *62*, 5881–5905.
28. Chen, Y.; Xu, J. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *J. Mach. Learn. Res.* **2016**, *17*, 882–938.

29. Hajek, B.; Wu, Y.; Xu, J. Submatrix localization via message passing. *arXiv* **2015**, arXiv:1510.09219.
30. Bubeck, S.; Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Found. Trends Mach. Learn.* **2012**, *5*, 1–122.
31. Cesa-Bianchi, N.; Lugosi, G. *Prediction, Learning, and Games*; Cambridge University Press: Cambridge, UK, 2006.
32. Lai, T.L.; Robbins, H. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* **1985**, *6*, 4–22.
33. Auer, P.; Cesa-Bianchi, N.; Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **2002**, *47*, 235–256.
34. Anthony, M.; Bartlett, P.L. *Neural Network Learning: Theoretical Foundations*; Cambridge University Press: Cambridge, UK, 1999.
35. Even-Dar, E.; Mannor, S.; Mansour, Y. PAC bounds for multi-armed bandit and Markov decision processes. In Proceedings of the Fifteenth Annual Conference on Computational Learning Theory, Sydney, Australia, 8–10 July 2002; Springer: Berlin, Germany, 2002; pp. 255–270.
36. Mannor, S.; Tsitsiklis, J.N. The sample complexity of exploration in the multi-armed bandit problem. *J. Mach. Learn. Res.* **2004**, *5*, 623–648.
37. Auer, P.; Cesa-Bianchi, N.; Freund, Y.; Schapire, R.E. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.* **2002**, *32*, 48–77.
38. Maillard, O.; Munos, R. Adaptive bandits: Towards the best history-dependent strategy. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 11–13 April 2011; pp. 570–578.
39. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and Regression Trees*; CRC Press: Boca Raton, FL, USA, 1984.
40. Hyvärinen, A.; Karhunen, J.; Oja, E. ICA by Minimization of Mutual Information. In *Independent Component Analysis*; John Wiley & Sons, Inc.: New York, NY, USA, 2002; pp. 221–227.
41. Janzing, D.; Mooij, J.; Zhang, K.; Lemeire, J.; Zscheischler, J.; Daniušis, P.; Steudel, B.; Schölkopf, B. Information-geometric approach to inferring causal directions. *Artif. Intell.* **2012**, *182*, 1–31.
42. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238.
43. Maes, F.; Collignon, A.; Vandermeulen, D.; Marchal, G.; Suetens, P. Multimodality image registration by maximization of mutual information. *IEEE Trans. Med. Imaging* **1997**, *16*, 187–198.
44. Wolpert, D.H.; Wolf, D.R. Estimating functions of probability distributions from a finite set of samples. *Phys. Rev. E* **1995**, *52*, 6841.
45. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *15*, 1191–1253.
46. Valiant, G.; Valiant, P. Estimating the unseen: An $n/\log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing, San Jose, CA, USA, 6–8 June 2011; pp. 685–694.
47. Jiao, J.; Venkat, K.; Han, Y.; Weissman, T. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory* **2015**, *61*, 2835–2885.