ISIT 2015 Tutorial: Information Theory and Machine Learning

Emmanuel Abbe* Martin Wainwright†

June 14, 2015

Abstract

We are in the midst of a data deluge, with an explosion in the volume and richness of data sets in fields including social networks, biology, natural language processing, and computer vision, among others. In all of these areas, machine learning has been extraordinarily successful in providing tools and practical algorithms for extracting information from massive data sets (e.g., genetics, multi-spectral imaging, Google and FaceBook). Despite this tremendous practical success, relatively less attention has been paid to fundamental limits and tradeoffs, and information theory has a crucial role to play in this context.

The goal of this tutorial is to demonstrate how information-theoretic techniques and concepts can be brought to bear on machine learning problems in unorthodox and fruitful ways. We discuss how any learning problem can be formalized in a Shannon-theoretic sense, albeit one that involves non-traditional notions of codewords and channels. This perspective allows information-theoretic tools—including information measures, Fano's inequality, random coding arguments, and so on—to be brought to bear on learning problems.

We illustrate this broad perspective with discussions of several learning problems, including sparse approximation, dimensionality reduction, graph recovery, clustering, and community detection. We emphasise recent results establishing the fundamental limits of graphical model learning and community detection. We also discuss the distinction between the learning-theoretic capacity when arbitrary "decoding" algorithms are allowed, and notions of computationally-constrained capacity. Finally, a number of open problems and conjectures at the interface of information theory and machine learning will be discussed.

*Program in Applied and Computational Mathematics, and Department of Electrical Engineering, Princeton University, Princeton, USA, eabbe@princeton.edu, www.princeton.edu/∼eabbe

†Departments of Electrical Engineering and Computer Science, and Department of Statistics, University of California at Berkeley, Berkeley, USA, wainwrig@berkeley.edu, http://www.cs.berkeley.edu/∼wainwrig.

Contents

1 Part I

1.1 Introduction [Slides 1–6]

- Standard books on concentration of measure include Ledoux [Led01] and Boucheron et al. [BLM13]; see also the sources from this morning's tutorial!

- The curse of dimensionality manifests itself in various ways, both computational and statistical. Roughly speaking, it refers to the fact that computational and statistical complexities for many problems explode exponentially in the dimension of the problem. See slide 36 for a concrete illustration in the context of non-parametric regression.

- There are various "no free lunch" theorems that show that, in the *absence of low-dimensional structure*, very little can be done in high-dimensional learning problems with limited samples. Standard examples of low-dimensional structure include sparsity (in vectors or matrices), rank constraints or eigendecay (in matrices), as well as additive decompositions and ridge models (for regression, density estimation). See the overview papers [NRWY12, Wai14b] for more details about high-dimensional statistics.

- There are various well-studied connections between information theory and statistics (e.g., see Cover and Thomas [CT91]) including the role of Kullback-Leibler divergence in hypothesis testing, relations between Fisher information and KL divergence, etc.

- Particularly relevant connections for this tutorial are the general view of a statistical estimation problem in terms of channel coding, and the intimate connections between metric entropy, Fano's inequality, and lower bounds on the minimax risk.

- In their classic paper, Kolmogorov and Tikhomirov [KT59] make connections between statistical estimation, metric entropy and the notion of channel capacity.

1.2 Graphical model selection

1.2.1 Background and past work [Slides 9–13]

- The model for binary random variables considered here is a heterogeneous form of Ising model [Isi25], in which the edge weights are allowed to be distinct. It is a very special case of a graphical model; see the books [KF10, WJ08] for more details on graphical models in machine learning.

- Chow and Liu [CL68] showed that the problem of finding the maximum likelihood tree is equivalent to solving a maximum weight spanning tree problem, which can be solved in polynomial-time. Karger and Srebro [KS01] show that this desirable property does not extend to graphs of treewidth larger than one. (By definition, an ordinary tree has treewidth one.)

- By the Markov properties of graphical models, detecting the absence of edges is equivalent to detecting conditional independence relationships among variables. There is a very large literature on the use of testing methods for graph selection (e.g., [KB07, SGS00, BMS13, ATHW12]).

- The idea of pseudolikelihood dates back to the work of Besag [Bes75, Bes77]. Csiszar and Talata [CT06] analyze a model selection method based on combining pseudolikelihood with a BIC penalty [Sch78].

- Meinshausen and Buhlmann [MB06] proposed the use of the Lasso (ℓ_1-regularized linear regression) as a method for performing neighborhood selection at a given node of a graphical model, and provided some high-dimensional consistency results allowing for the number of nodes p to be larger than the sample size n. Zhao and Yu [ZY06] prove consistency under milder conditions, whereas Wainwright [Wai09b] provides sharp lower and upper bounds on the performance of ℓ_1-relaxation for variable selection, as well as various information-theoretic lower bounds [Wai09a]. See also the papers [FRG09, AT10] for related results.

- For binary graphical models, Ravikumar et al. [RWL10] analyzed the performance of ℓ_1-regularized logistic regression, corresponding to a form of neighborhood selection. The performance of ℓ_1-relaxations is well-known to depend on certain "incoherence" conditions; Bento and Montanari [BM09] proved that these conditions are violated with high probability at some point above the phase transition level for the Ising model.

- A variety of other simple algorithms have been explored for selecting discrete graphical models. For instance, Bresler et al. [BMS13] analyze a simple thresholding and testing algorithm for discrete graphical models, with emphasis on the case of bounded degree graphs. Netrapalli et al. [NBSS10] analyzes a greedy algorithm, whereas Anandkumar et al. [ATHW12] analyze the performance of a local testing method under various notions of local separation. Bresler [Bre14] proposes an algorithm for learning arbitrary Ising models (without incoherence conditions) that is efficient on bounded degree graphs.

1.2.2 Role of the ratio $n/(d^2 \log p)$

- Slide 14: The data used to fit this network consisted of the voting records of the US Senate over the period 2006—2008.

- Slide 16: These curves were generated by applying the ℓ_1-regularized LR method [RWL10] to simulated data drawn from star-shaped graphs on p nodes with a maximum degree $d = \lceil 0.1 \rceil$, obtained at the hub node. The plots shows the probability of the method correctly recovering the graph (all edges correct, no false inclusions) versus the sample size n.

- Slide 17: This slide shows the same probability of correct recovery, now plotted versus the rescaled sample size $n/(d^2 \log p)$ where d is the maximum degree, and p is the number of nodes. It shows empirically that this ratio is the right way of measuring the capacity in a broad class of graph recovery problems. The theory on Slide 18 provides confirmation that this ratio is the correct order parameter.

1.3 Sparse PCA

1.3.1 Basics [Slides 22–24]

- Principal component analysis is a widely used technique for dimension reduction and data compression; see the books [Jol04, And84] for furthe details.

- Johnstone [Joh01] discusses the inconsistency of classical PCA in the high-dimensional setting, and considers various forms of spiked covariance models. For a standard spiked model (without sparsity), there is an interesting asymptotic transition that occurs as a function of p/n. In particular, if $p/n \to \alpha$, then there is a threshold SNR $\nu = \nu(\alpha)$ below which the maximal eigenvector of the sample covariance matrix is orthogonal to the population eigenvector. Essentially, the high-dimensional noise (due to having $p \asymp n$ directions) accumulates too quickly, and swamps the signal.

- Slides 23–24: These are instances of what are known as "eigenfaces". They can be used to perform forms of face recognition. The sparse eigenfaces were approximated using a truncated form of the power method for computing eigenvectors.

1.3.2 Diagonal thresholding and optimal algorithm [Slides 25–28]

- The spiked covariance model was first introduced by Johnstone and Lu [JL09a], who also proposed diagonal thresholding (DT) as a pre-processing step for a multistage procedure.

- Amini and Wainwright [AW09a] proved matching upper and lower bounds on the DT method for variable selection, proving that $n/(k^2 \log p)$ is the correct order parameter. They also used an information-theoretic argument—in particular, one based on the Fano method—to prove that the transition of an optimal method occurs as a function of $n/(k \log p)$.

1.3.3 SDP relaxation and computational barriers [Slides 29–34]

- Slide 29: The "lifting procedure" described here is a standard one for deriving SDP relaxations of nonconvex programs. Its application to sparse PCA is due to d'Aspremont et al. [dEJL07].

- Amini and Wainwright [AW09a] showed that the SDP relaxation is always at least as good as the DT method. Moreover, they proved a conditional guarantee: namely, that if a rank one solution exists, then the SDP method has a recovery threshold of order $n/(k \log p)$. Empirically, they observed that a rank one solution exists for sparsity $k \ll \log p$; other authors have proved that rank one solutions do not exist in the "hard regime".

- Berthet and Rigollet [BR13] prove that conditioned on the (conjectured) average-case hardness of the planted k-clique problem, there is a gap between the classical and polynomial-constrained minimax rates of detection. Since this work, other authors [MW13] have proved related gaps, using different reductions also based on planted clique.

1.4 Structured non-parametric regression

1.4.1 Basics [Slides 35—37]

- Non-parametric regression is widely studied in statistics and machine learning; see the books [GKKW02, Tsy09, vdG00, EL07, GKKW02, HMSW04, Was06] for more background.
- The notion of minimax risk plays a central role in mathematical statistics, and there is a very large literature devoted to techniques for upper and lower bounding it (e.g., [LC73, Bir83, Bir87, Tsy09, HI90, Has78, Yu93, YB99, Gun11]). In general, the primary interest is in obtaining upper and lower bounds that capture the dependence on sample size n, dimension d, as well as other structural parameters of the problem (e.g., smoothness, sparsity etc.).
- In the classical formulation of minimax theory, the infimum over all estimators is completely unconstrained, and hence allows for estimators that may have prohibitive computational or memory requirements. A more recent line of work (e.g., see the paper [Wai14a] and references therein) is exploring the notion of constrained statistical minimax, in which this infimum is further constrained (based on computational, storage, or privacy constraints). The computational gap in the sparse PCA problem (slide 32) can be understood as an instance in which there is a substantial gap between the classical and computationally-constrained minimax rates.

1.4.2 Metric entropy [Slides 38—40]

- Slides 38, 39: the notion of metric entropy emerged from the Russian school, with major contributions by Kolmogorov and his collaborators; see the seminal paper [KT59] for a beautiful overview. It is a central concept in the area of approximation theory; see the books [Pin85, DL93] for further details.
- Slide 40: See the paper [KT59] for a more detailed discussion of this case, as well as many more examples (e.g., involving higher order smoothness constraints).

1.4.3 Reduction and master equation [Slides 41–42]

- Slide 41: This standard reduction is the first step in any of the Le Cam, Assouad or Fano methods for lower bounding the classical minimax risk. See the overview paper by Yu [Yu93] for a comparison of these three approaches, as well as Chapter 15 from the book in preparation [Wai15].
- Slide 42: There are many different forms of master equations for obtaining lower bounds on minimax rates; these involve either quantities like the metric entropy, or in certain pairwise cases, related objects such as moduli of continuity. Le Cam [LC73] provided the first, involving the Hellinger distance; see also Donoho and Liu [DL91] for a broader framework based on this idea. The master equation in Slide 42 arises from a special case of a more general framework due to Yang and Barron [YB99]. (For our specific regression model, the error metric $\|\cdot\|_n$ is the same as the KL divergence up to constant pre-factors, which leads to this simplified matching condition.)

1.4.4 Examples [Slides 43–45]

- Slide 43: The problem of sparse linear regression has been studied intensively in the context of both compressed sensing, and also in statistics and machine learning. Compressed sensing shares parallels with coding theory, in that the choice of the vectors x_i is under the user's control (and are typically chosen to induce favorable properties). In machine learning applications, the covariate vectors x_i are a design parameter, and it is desirable to have results that allow for substantial correlations between the elements of x_i—a condition excluded by incoherence or restricted isometry conditions.

- A large body of work has focused on the performance of ℓ_1-relaxations (e.g., [Nem00, GR04, CT05, CT07, Don06a, Don06b, BRT09, vdGB09]), and their behavior is now very well-understood.

- Raskutti et al. [RWY11] prove upper and lower bounds on the minimax error in sparse regression over ℓ_q-"balls", over the range $q \in [0, 1]$—in particular, sets of the form

$$\mathbb{B}_q(R_q) = \big\{\theta \in \mathbb{R}^p \mid \sum_{j=1}^{p} |\theta_j|^q \leq R_q\big\}.$$

  This analysis, combined with known results on the performance of ℓ_1-relaxation over ℓ_q-balls [NRWY12], shows that ℓ_1-methods achieve minimax rates for the ℓ_2-error $\|\widehat{\theta} - \theta^*\|_2$.

- In sharp contrast, for the prediction error $\|X(\widehat{\theta} - \theta^*)\|_2/\sqrt{n}$, it is known that the Lasso or any related ℓ_1-method will not achieve the minimax rate unless the matrix X satisfies restrictive conditions like the RE condition. These conditions are not necessary for an optimal (exponential-time) method, so that there is a gap between ℓ_1-performance and optimal performance. This gap turns out to be fundamental, as can be made precise by a rigorous complexity-theoretic analysis [ZWJ14, ZWJ15].

- Slide 44: The metric entropy scaling stated in this slide is proved in the paper [KT59].

- There are a large number of computationally efficient methods that achieve the $n^{-\frac{2\alpha}{2\alpha+1}}$ minimax rate for α-smooth regression, including kernel estimators with appropriately chosen bandwidth [Was06] methods based on orthogonal series expansion [Tsy09], as well as M-estimators based on reproducing kernel Hilbert spaces [RWY14].

- Slide 45: There are a broad range of structured models for non-parametric regression. Minimax rates associated with additive models were derived by Stone [Sto85]; see also the papers [HT86, BHT89]. Various authors have studied variants of the sparse additive model (e.g., [MvdGB09, RLLW09, KY10]). The minimax rates sketched out here were proved in detail by Raskutti et al. [RWY12].

2 Part II

2.1 Introduction [Slides 1-9]

Inverse problems on graphs. A large variety of machine learning and data-mining problems are about inferring global properties on a collection of agents by observing local noisy interactions of these agents.

[Slide 2]. A prominent example is the problem of detecting communities in social and biological networks [For10]. Slide 1 shows two examples of social networks, a user's Facebook network[1] and a blog network[2]. In the Facebook network, vertices represent users and edges represent friendships between users. In the blog network, vertices represent blogs and edges represent hyperlinks between blogs (ignoring directions). In both cases, a problem of major interest is to identify groups of nodes that are "alike", i.e., communities (further discussion on these notions will come later on). Considering that each node has some hidden community attribute, and assuming that edges are placed placed depending on these attributes, in a noisy fashion (e.g., two people in the same community may not always connect, and conversely, two people in different communities may sometimes connect), this problem is an inverse problem where the edges are observed and the node attributes are unknown.

[Slide 3] A different example concerns image segmentation in computer vision. Slide 2 represents segments of a lizard picture[3]. An approach pioneered by [SM97] to extract image segments is to build a graph of similarities of the pixels (or group of pixels). Simplifying things, define a graph where nodes are pixels and connect pixels with some measure of similarity which may depend on both their proximity and their colors/intensities (see the cup picture[4]). The hidden attributes of the nodes are in this case the segments that they belong to, and the edges of the graph provide noisy measure that depend upon these. Hence we have again an inverse problem from edge to node variables.

[Slide 4] The previous approach uses what is called a "graph of similarity," which connects elements in data sets based on their pairwise relationships, and from which one desires to reconstruct clusters of similar nodes. This is a broadly employed approach in machine learning and data-mining. Slide 5 provides[5] three additional examples: (i) information retrieval in text documents, where words may be connected based on co-occurrences, (ii) unsupervised image classification, where images are connected based on joint features, (iii) products categorizations, where books or movies are connected based on similar ratings of users, or conversely, where users are connected based on similar ratings of products. In all of these, the graph provides local noisy measurements of hidden variables about the nodes, and one wishes to recover these variables, or their similarity classes. Identifying

[1] Taken from https://griffsgraphs.wordpress.com/2012/07/02/a-facebook-network/

[2] Taken from http://allthingsgraphed.com/2014/10/09/visualizing-political-polarization/

[3] Taken from http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/

[4] Taken from http://scikit-image.org/docs/dev/auto_examples

[5] Illustrative images taken from Google images.

these clusters can then be used to mind data, to recognize images of a same type, to cluster users for more accurate recommendations, to unveil biological functionality of groups of proteins (PPI networks), to improve medical diagnosis based on similar patient records etc.

[Slide 6] The common model for all these examples is that a network is generated from hidden variables at the nodes and one wishes to solve the inverse problem. At the end of this note, we will see that dimensionality reduction problems (e.g., sparse-PCA) can also be related to such inverse problems. Note that this is somewhat dual to the problem of learning graphical models, where the node variables are observed, and where one wishes to recover the underlying edges that capture the node dependencies.

[Slide 7] Are graph-based codes another such example? Note quite, but not too far either. The problem of decoding a code is also an inverse problems, where noisy observations are made based on groups of unknown node variables. In fact, the channel produces noisy outputs of groups of the node variables (the information bits). The main difference with machine learning problems is that the code is a design parameter in communications, and can take some quite sophisticated forms, that may not be local (think of a random or polar code). Even an LDGM code that takes local XORs of the bits, though typically not pairwise, is still very different since one also constraints the left node degrees. Traditional codes are hence not natural models for ML applications. On the other hand, once admitted that the code emerging from ML applications is not a design parameter and is somewhat *unorthdox*, the channel perspective and decoding problems remain effective. We hence ask the following:

- What are the relevant channels and codes in the ML applications?
- What are the recovery requirements?
- Are there fundamental limits to the decoding problems in such models?

We next focus on community detection and clustering which encompass most of the applications previously discussed.

2.2 Community detection and clustering [Slides 10-11]

In community detection, one wishes to recover similarity classes of the node attributes. For the purpose of this tutorial, we will consider communities and clusters as being equivalent notions. These are often interchanged in the literature, though clusters may sometimes refer more specifically to groups of nodes having denser intra-connectivity, or higher proximity in some metric, whereas communities may also represent more abstract similarity relationships, which can be disassortative (e.g., with more connections across the groups). In this tutorial, we will use 'community detection' or 'clustering' for same purposes.

Detecting communities (or clusters) in graphs is a fundamental problem in networks, computer science and machine learning. This applies to a large variety of complex networks (e.g., social and biological networks) as well as to data sets engineered as networks via similarly graphs, where one often attempts to get a first impression on the data by trying to identify groups with similar behavior. In particular, finding communities allows one to find like-minded people in

social networks [GN02, NWS], to improve recommendation systems [LSY03, XWZ+14], to segment or classify images [SM97, SHB07], to detect protein complexes [CY06, MPN+99], to find genetically related sub-populations [PSD00, JTZ04], or discover new tumor subclasses [SPT+01].

2.3 Exporting Shannon's program [Slides 12-13]

Community detection is not a new problem. It has been around for at least two/three decades, and it exploded more recently with the emergence of social networks. The problem is notoriously hard. Defining what is a good clustering can be subject to debates in applications. However, even if one uses a model with groundtruth to set this question, the problem of recovering the communities is typically hard. As well see, even the most basic problems such as min-bisection are NP-hard [DF89, BCLS87, CK99, BS04]. Consequently, myriads of heuristics have been proposed for community detection methods, driven mainly efficiency criteria. This is in particular a well known challenge for Big Data problems where one cannot manually determine the quality of the clusters [RDRI+14], and where computational barriers take place.

The goal in this tutorial is somehow to shift the focus temporarily from algorithms to information theory. We mention 'temporarily' because eventually the goal is to obtain efficient algorithms. However, we plan to follow the recipe of information theory in data transmission: to understand how to devise efficient and robust algorithms (i.e., that succeed in the most challenging regimes), we will first look for an understanding of the *fundamental limits* of community detection problems. This will hopefully allow us to set benchmarks and line-of-sights for algorithms, similarly to Shannon's coding theorem which has driven coding theory (and part of the communication industry) for many years. Such a line-of-sight has been missing in community detection, but has started to appear more recently, in particular for the stochastic block model [Co10, DKMZ11, Mas14, MNS14, ABH14, MNSb, YC14, AS15a].

2.4 The Stochastic Block Model [Slides 14-15]

[Slide 14] The stochastic block model (SBM) is a canonical model for community detection. It is one of the most popular network models exhibiting community structures [HLL83, WBB76, FMW85, WW87, BC09, KN11]. The model was first proposed in the 80s [HLL83] and received significant attention in the mathematics and computer science literature [BCLS87, DF89, Bop87, JS98, CK99, CI01, McS01], as well as in the statistics and machine learning literature [SN97, BC09, RCY11, CWA12]. The SBM puts a distribution on n-vertices graphs with a hidden (or planted) partition of the nodes into k communities. Denoting by p_i, $i \in [k]$, the relative size of each community, and assuming that pairs of nodes in communities i and j connects independently with probability $Q_{i,j}$, the SBM can be defined by the triplet (n, p, Q), where p is a probability vector of dimension k and Q a $k \times k$ symmetric matrix with entries in $[0, 1]$.

The success of the SBM relies on the fact that it is overall a good model. This does not mean that it is 100% realist, but it is an insightful model for community detection, and can be realist with slight extensions. In a sense, it plays a similar role to the DMC in communications.

[Slide 15 & 36] The SBM recently came back at the center of the attention at both the practical level, due to extensions allowing overlapping communities [ABFX08] that have proved to fit well real data sets in massive networks [GB13], and at the theoretical level due to new phase transition phenomena discovered for the two-community case [Co10, DKMZ11, Mas14, MNS14, ABH14, MNSb, YC14, AS15a]. To discuss these phenomena, we need to first introduce the figure of merits:

- **Weak recovery** (also called detection). This only requires the algorithm to output a partition of the nodes which is positively correlated with the true partition (whp[6]). Note that weak recovery is relevant in the fully symmetric case where all nodes have identical average degree,[7] since otherwise weak recovery can be trivially solved. If the model is perfectly symmetric, like the SBM with two equally-sized clusters having the same connectivity parameters, then weak recovery is non-trivial. Full symmetry may not be representative of reality, but it sets analytical and algorithmic challenges. The weak-recovery threshold for two symmetric communities was achieved efficiently in [Mas14, MNS14], settling a conjecture established in [DKMZ11]. The case with more than two communities remains open.

- **Partial recovery.** One may ask for the finer question of *how much* can be recovered about the communities. For a given set of parameters of the block model, finding the proportion of nodes (as a function of p and Q) that can be correctly recovered (whp) is an open problem. Partial results were obtained in [MNSa] for two-symmetric communities, and the a general result for large degrees is given in [AS15a].

- **Almost exact recovery** One may also consider the special case of partial recovery where only an $o(n)$ fraction of nodes is allowed to be mis-classified (whp), called almost exact recovery or weak consistency, but no sharp phase transition is to be expected for this requirement.

- **Exact recovery** (also called recovery or strong consistency.) Finally, one may ask for the regimes for which an algorithm can recover the entire clusters (whp). This is non-trivial for both symmetric and asymmetric parameters. One can also study "partial-exact-recovery," namely, which communities can be exactly recovered. While exact recovery has been the main focus in the literature for the past decades, the phase transition for exact recovery was only obtained last year for the case of two symmetric communities [ABH14, MNSb]. The general case has been solved in [AS15a].

2.5 The two-symmetric case (2-SBM) [Slides 16-24]

We refer to [ABH14] for the results presented in slides 16-24.

The main result states that recovery is possible in the regime $p = a\log(n)/n$, $q = b\log(n)/n$, $a, b > 0$, if and only if

$$\sqrt{a} - \sqrt{b} \geq 2 \tag{1}$$

[6] whp means with high probablity, i.e., with probability $1 - o_n(1)$ when the number of vertices diverges.

[7] At least for the case for communities having linear size. One may otherwise define stronger notions of weak recovery that apply to non-symmetric cases.

with an efficient algorithms achieving the threshold. Moreover, this is the bottleneck regime, and other regimes of p and q can be deduced from the above, as long as $p \nsim q$ (i.e., as long as p and q are not of the exact same order).

2.6 The general SBM [Slides 25-40]

We summarize below slides 25-40 and refer to [AS15a] for the details (and [AS15b] for the case where the parameters are unknown).

The main results in [AS15a] are for partial, almost exact and exact recovery, respectively in the constant, $\omega(1)$ and $\log(n)$ degree regime.

I. Partial and almost exact recovery in the general SBM. The first result of [AS15a] concerns the regime where the connectivity matrix scales as Q/n for a positive symmetric matrix Q (i.e., the node average degree is constant). The following notion of SNR is introduced[8]

$$\mathrm{SNR} = |\lambda_{\min}|^2/\lambda_{\max} \tag{2}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are respectively the smallest and largest eigenvalue of $\mathrm{diag}(p)Q$.

The algorithm `Sphere-comparison` is proposed that solves partial recovery with exponential accuracy and quasi-linear complexity when the SNR diverges, solving in particular almost exact recovery.

The following is an important consequence of the main theorem in [AS15a], as it shows that `Sphere-comparison` achieves almost exact recovery when the entries of Q are scaled.

Theorem 1. *[AS15a] For any $k \in \mathbb{Z}$, $p \in (0,1)^k$ with $|p| = 1$, and symmetric matrix Q with no two rows equal, there exists $\epsilon(\delta) = O(1/\ln(\delta))$ such that for all sufficiently large δ there exists an algorithm (`Sphere-comparison`) that detects communities in graphs drawn from SBM$(n, p, \delta Q)$ with accuracy $1 - e^{-\Omega(\delta)}$ and complexity $O_n(n^{1+\epsilon(\delta)})$.*

For k symmetric clusters, SNR reduces to $\frac{(a-b)^2}{k(a+(k-1)b)}$, and [AS15a] shows that if the SNR diverges, then almost exact recovery is solvable efficiently. Moreover, the SNR must diverge to ensure almost exact recovery in the symmetric case.

II. Exact recovery in the general SBM. The second result in [AS15a] is for the regime where the connectivity matrix scales as $\log(n)Q/n$, Q fixed, where it is shown that exact recovery has a sharp threshold characterized by the divergence function

$$D_+(f,g) = \max_{t\in[0,1]} \sum_{x\in[k]} \left(tf(x) + (1-t)g(x) - f(x)^t g(x)^{1-t}\right),$$

named the CH-divergence in [AS15a]. Specifically, if all pairs of columns in $\mathrm{diag}(p)Q$ are at D_+-distance at least 1 from each other, then exact recovery is solvable in the general

[8] Note that this in a sense the "worst-case" notion of SNR, which ensures that all of the communities can be separated (when amplified); one could consider other ratios of the kind $|\lambda_j|^2/\lambda_{\max}$, for subsequent eigenvalues ($j = 2, 3, \dots$), if interested in separating only subset of the communities.

SBM. This provides in particular an operational meaning to a new divergence function analog to the KL-divergence in the channel coding theorem (see Section 2.3 in [AS15a]). Moreover, an algorithm (`Degree-profiling`) is developed that solves exact recovery down to the D_+ limit in quasi-linear time, showing that exact recovery has no informational to computational gap (as opposed to the conjectures made for detection with more than 4 communities [DKMZ11]). The following gives a more general statement characterizing which subset of communities can be extracted.

Theorem 2. *[AS15a]* **(i)** *Exact recovery is solvable in the stochastic block model $\mathbb{G}_2(n,p,Q)$ for a partition $[k]=\sqcup_{s=1}^t A_s$ if and only if for all i and j in different subsets of the partition,*[9]

$$D_+((PQ)_i,(PQ)_j)\geq 1, \tag{3}$$

In particular, exact recovery is information-theoretically solvable in SBM$(n,p,Q\log(n)/n)$ if and only if $\min_{i,j\in[k],i\neq j} D_+((PQ)_i||(PQ)_j)\geq 1$.
(ii) *The* `Degree-profiling` *algorithm (see [AS15a]) recovers the finest partition that can be recovered with probability $1-o_n(1)$ and runs in $o(n^{1+\epsilon})$ time for all $\epsilon>0$. In particular, exact recovery is efficiently solvable whenever it is information-theoretically solvable.*

In summary, exact or almost exact recovery is closed for the general SBM (and detection is closed for 2 symmetric communities). However this is for the case where the parameters of the SBM are assumed to be known, and with linear-size communities.

The case where the parameters are unknown was recently solved in [AS15b], showing that the parameters (including the number of communities) in the general SBM with linear size communities can be efficiently learned.

2.7 An example with real data [Slide 41-42]

We refer to [AS15b] for an example on a real data set, detecting communities in the political blog network of Adamic and Glance [AG05]. In particular, the algorithm of [AS15b] is shown to achieve the state-of-the-art for this data set.

2.8 Open problems on community detection [Slide 44-53]

Below we present some open problems that have information theoretic flavor.

2.8.1 Stochastic block model

A. Recovery. [Slide 44] The CH-divergence is shown to be the fundamental limit for recovery in [AS15a] when the relative size of communities is lower-bounded, i.e., linear-size communities. One may wonder how general is the connection between f-divergences and clustering problems. We will see later other types of block models for which such a connection may also take place. Here we simply mention that recovery may also be studied in other regimes. First, the results of [AS15a] should extend to a slowly growing number of communities, up to logarithmic orders, but there may be new phenomena taking place

[9] The entries of Q are assumed to be non-zero.

in the logarithmic regime or beyond. In [YC14], the symmetric SBM with k communities is considered in regimes where p, q and k scale polynomially with n (see also [CSX12]). Focusing on coarse scalings of the parameters, interesting regimes are identified where community recovery is either solvable information-theoretically or computationally. One may further look at the phase transitions for these regimes.

A. Detection and broadcasting problems. [Slide 45-46] The converse result in [MNS12] shows that in the regime $p = a/n$, $q = b/n$, detection is impossible if $(a-b)^2 \leq 2(a+b)$. This is obtained by a reduction to the broadcasting problem on tree [EKPS00], which is a nice information-theory problem. Consider a bit which is broadcasted over a tree with independent BSCs on each branch. The number of children is governed by the so-called offspring distribution of the tree. If the tree is viewed as the local neighborhood of a node in the SBM, this is a Poisson distribution of mean $c = (a+b)/2$, and the BSCs have noise parameter $\varepsilon = b/(a+b)$. One can now ask the question of when (in terms of the offspring distribution and the noise parameter) the bit can be *detected* given the leaves values at large depth. Note that this is an unorthodox broadcasting problem (compared to traditional information theoretic broadcasting problems), since one wishes only to recover the bit with a probability that is away from $1/2$ rather than close to 1. In [EKPS00], this problem is solved for a broad class of offspring distributions, showing that detection is possible if and only if $c > 1/(1-2\varepsilon)^2$. This gives $(a-b)^2 > 2(a+b)$ if the parameters are chosen as above.

Interestingly, the general extension of the broadcasting problem on tree with non-binary variables is open. It is conjectured that gaps between information-theoretic and computational reconstructions take place for larger alphabets. This resonate with the following conjecture for multiple communities made in [DKMZ11], for the symmetric model with k communities (probability a/n inside the clusters and b/n across).

Conjecture 1. *For the symmetric k-SBM(n, a, b), there exists c_k s.t.*
(1) If SNR $< c_k$*, then detection cannot be solved,*
(2) If $c_k <$ SNR < 1*, then detection can be solved information-theoretically but not efficiently,*
(3) If SNR > 1*, then detection can be solved efficiently. Moreover $c_k = 1$ for $k \in \{2, 3, 4\}$ and $c_k < 1$ for $k \geq 5$.*

C. Partial-recovery and the SNR-distortion curve. [Slide 47] Beyond detection, and before almost exact recovery, one may ask *how much* can be recovered about the communities at finite SNR. As shown in slide 49, an SNR-distortion curve is conjectured to take place, which extends the above conjecture. There are three partial results about this problem:
(1) [MNSa] gives the answer for large enough SNR in the case of two symmetric communities (using the broadcasting on tree problem);
(2) [AS15a] gives the behaviour for large SNR in the case of the general SBM;
(3) [DAM15] gives the behaviour for finite SNR but when $(a-b)^2$ and $a+b$ are both diverging (keeping their ratio constant).

The general expression remains unkonwn. Techniques from information theory are likely to be insightful here.

2.8.2 Other block models

[Slide 48-52] There is a broad set of other block models, extending from the SBM. Three important extensions for applications are models with corrected-degrees [KN11], overlapping communities [ABFX08] and labelled edges [HLM12, XLM14]. Various partial results have been obtained for these models, however the information-theoretic tradeoffs discussed previously for recovery, detection and partial recovery are yet to be resolved.

A natural model for many real networks is the OSBM described for example in [AS15a], where each node in the graph has a node-profile consisting of s bits (for, say, s attributes) and where pairs of nodes connect in proportion to the number of joint attributes. One can look at this as an SBM with 2^s communities, but then the complexity of the methods in [AS15a] scale terribly. Hence a different view is need for large s.

An interesting and simple variant of the 2-SBM is the 2-CBM. The model was defined in [AM13, ABBS14a] and is a special case of [HLM12]. It is also related to many other models [BBC04, KPSS10, AM, HG13, CHG, CG14, ABBS14b, GRSY14, PM14].

The 2-CBM(n, p, ε) is defined as follows. First a standard ER graph is drawn with edge probability p. The nodes are split into two balanced clusters independently, and edges of the ER graphs are colored in blue inside the clusters and red outside the clusters. Finally, each color is flipped with probability ε. An SDP relaxation was proposed in [ABBS14a] for the CBM, with a performance gap having roughly a factor 2. This gap was recently closed in [BH14, Ban15]. It was also shown in [YP14] that a spectral algorithm achieves the threshold. The detection and recovery thresholds are now known for this model [BH14, CRV15, SKLZ15]. Note that SDP relaxations for block models were also studied in [YC14, AL14, GV14].

Establishing the SNR-distortion curve for 2-CBM is open but may be reachable. The recovery for k symmetric communities is also known [BH14], but not the asymmetric case. In particular, a CH-divergence may take place again as the fundamental limit for recovery in the general CBM (open). Similar conjectures are expected to take place for detection.

A model with a single planted community was recently studied in [Mon15]. In such a case, the size of the planted community is much smaller than the ambient graph. If the edge probability is 1 in the planted community, it becomes planted clique (see [DM13] for example). For this model, [Mon15] shows that a detection gap is also expected to take place, in even a coarser regime than for the SBM.

2.8.3 Beyond block models

[Slide 53] Graphons [Lov12] are natural extension of SBMs. A graphon is defined with a continuous kernel $w : [0, 1]^2 \to [0, 1]$. Each vertex v in the graph is assigned a random uniform number x_v in [0,1], and edges are placed independently conditioning on the vertex labels from the kernel w. How SBM can approximate graphons (in the spirit of piecewise constant approximations of functions) under some regularity assumptions was studied in [CWA12, ACC13]. These papers (as well as [BCS15]) study in particular the problem of "estimating" graphons. The recovery problems for graphons, i.e., how node attributes may be approximated, represent interesting problems. Note that there is a more general perspective of graphons as limits of graphs related to Szemerédi regularity lemma, we refer to [Lov12] for more details.

2.9 Graphical channels [Slide 54]

Graphical channels were defined in [AM13, AM] as a class of channels inspired by inference problems on graphs. They can be seen as memoryless channels encoded with an unorthodox code which only takes k information symbols at a time. In other words, the tanner graph of the code is a hypergraph with bounded edge-order. All examples discussed in this tutorial can be casted as graphical channels (including sparse-PCA as next discussed).

It is natural to ask whether general properties of graphical channels can be obtained. In [AM], we consider the case of uniform models for the graph, and discuss whether the normalized mutual information $\frac{1}{n}I(X;Y)$ admits a limit. While this seems true in a general, it is surprisingly hard to prove. We could only obtain the result for certain cases of kernels (which includes most symmetric kernels), showing a sub-additivity property of the mutual information. A different approach may exist to simplify such results. In addition, the value of the limit is not found in [AM], which an interesting open problem. In terms of decoding nodes reliably for graphical channels, it is possible that a CH-divergence limit takes place for general cases. This is also an open problem.

2.10 Connection clustering/sparse-PCA [Slide 55-56]

Consider the following spiked Wigner model (see [JL09b] for spiked covariance matrix models)

$$Y_\lambda = \sqrt{\frac{\lambda}{n}} XX^t + Z, \tag{4}$$

where Z is Wigner (i.e., symmetric with i.i.d. standard Gaussian entries). In sparse-PCA, X is i.i.d. Bernoulli(ε), i.e., there are about $k = \varepsilon n$ one's (one may consider other sparsity model), and the problem is to recover X from Y_λ. Two references for information-theoretic results are [AW09b, DM14].

How does this relate to the SBM, or more generally to clustering problems? The expected adjacency matrix of the 2-SBM is also rank one, so the 2-SBM is also a perturbation of a rank one matrix. More generally, the k-SBM is a rank k perturbation problem. However, the 2-SBM has two main differences with the sparse-PCA problem above: (1) the SBM is a purely binary model, i.e., the perturbation is also binary, whereas the spiked Wigner models is a continuous perturbation, (2) the expected adjacency matrix in the SBM has "blocks" that are balanced, where as XX^t gets sparser for a binary vector X.

The second difference above can be easily adapted. Consider the same model as in (4) but assume that X is instead i.i.d. Radamacher(1/2). Then XX^t is exactly like the expected adjacency matrix of the SBM (calling the community variables $+1$ and -1). Of course, the new model — a *blocked-spiked Wigner model* — is still different than the SBM since it is a continuous noise model. Is it really different though? Note that wstablishing an equivalence between the SBM and the blocked-spiked Wigner model would interesting to connect the problems and allow the exportation of methods specific to Gaussian noise models.

In [DAM15], it is shown that the equivalence holds when the SNR of the 2-SBM tends to the SNR parameter λ in the blocked-spiked Wigner model, *but* when the average degrees of the 2-SBM diverge. Using this connection, [DAM15] is able to compute the normalized mutual information $I(X;G)/n$ and normalized MMSE between the nodes variables and the

graph, with a single-letter formula. The latter exploits in particular the I-MMSE formula [GSV05] which holds for Gaussian noise models. It further shows how to achieve the optimal MMSE bound in the 2-SBM for that regime with an efficient AMP algorithm. The techniques are similar to [DM14]. We refer to [DAM15] for the result presented in Slide 58.

2.11 Conclusion [Slide 57]

Community detection couples naturally with the channel view of information theory and more specifically with:

- graph-based codes,
- f-divergences,
- broadcasting problems,
- I-MMSE,

in particular, with unorthodox versions of these. Much more of these connections are expected to take place, involving existing information-theoretic tools or requiring the development of new ones.

More generally, the problem of inferring global similarity classes in data sets from noisy local interactions is at the center of many problems in machine learning, and an information-theoretic view of these problems seems both needed and powerful.

2.12 Selected publications

Main results discussed: [DKMZ11], [MNS12], [Mas14], [MNS14], [ABH14], [MNSb], [ABBS14a], [AS15a], [AS15b], [DM14], [DAM15].

References

[ABBS14a] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer, *Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery*, IEEE Transactions on Network Science and Engineering **1** (2014), no. 1. 13, 15

[ABBS14b] E. Abbe, A.S. Bandeira, A. Bracher, and A. Singer, *Linear inverse problems on Erdös-Rényi graphs: Information-theoretic limits and efficient recovery*, Information Theory (ISIT), 2014 IEEE International Symposium on, June 2014, pp. 1251–1255. 13

[ABFX08] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, *Mixed membership stochastic blockmodels*, J. Mach. Learn. Res. **9** (2008), 1981–2014. 9, 13

[ABH14] E. Abbe, A. S. Bandeira, and G. Hall, *Exact recovery in the stochastic block model*, Available at ArXiv:1405.3267. (2014). 8, 9, 15

[ACC13] E. Airoldi, T. Costa, and S. Chan, *Stochastic blockmodel approximation of a graphon: Theory and consistent estimation*, arXiv:1311.1731 (2013). 13

[AG05] L. Adamic and N. Glance, *The political blogosphere and the 2004 u.s. election: Divided they blog*, Proceedings of the 3rd International Workshop on Link Discovery (New York, NY, USA), LinkKDD '05, ACM, 2005, pp. 36–43. 11

[AL14] A. Amini and E. Levina, *On semidefinite relaxations for the block model*, arXiv:1406.5647 (2014). 13

[AM] E. Abbe and A. Montanari, *Conditional random fields, planted constraint satisfaction and entropy concentration*, To appear in the journal *Theory of Computing*, available at arXiv:1305.4274v2. 13, 14

[AM13] _____, *Conditional random fields, planted constraint satisfaction and entropy concentration*, Proc. of RANDOM (Berkeley), August 2013, pp. 332–346. 13, 14

[And84] T. W. Anderson, *An introduction to multivariate statistical analysis*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York, 1984. 3

[ANP05] D. Achlioptas, A. Naor, and Y. Peres, *Rigorous Location of Phase Transitions in Hard Optimization Problems*, Nature **435** (2005), 759–764.

[AS15a] E. Abbe and C. Sandon, *Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms*, arXiv:1503.00609 (2015). 8, 9, 10, 11, 12, 13, 15

[AS15b] _____, *Recovering communities in the general stochastic block model without knowing the parameters*, arXiv:1506.03729 (2015). 10, 11, 15

[AT10] M. Akcakaya and V. Tarokh, *Shannon theoretic limits on noisy compressive sampling*, IEEE Trans. Info Theory **56** (2010), no. 1, 492–504. 2

[ATHW12] A. Anandkumar, V. Y. F. Tan, F. Huang, and A. S. Willsky, *High-dimensional structure learning of Ising models: Local separation criterion*, Annals of Statistics **40** (2012), no. 3, 1346–1375. 1, 2

[AW09a] A. A. Amini and M. J. Wainwright, *High-dimensional analysis of semdefinite relaxations for sparse principal component analysis*, Annals of Statistics **5B** (2009), 2877–2921. 3

[AW09b] Arash A Amini and Martin J Wainwright, *High-dimensional analysis of semidefinite relaxations for sparse principal components*, The Annals of Statistics **37** (2009), no. 5B, 2877–2921. 14

[Ban15] A. S. Bandeira, *Random laplacian matrices and convex relaxations*, arXiv:1504.03987 (2015). 13

[BBC04] N. Bansal, A. Blum, and S. Chawla, *Correlation clustering*, Mach. Learn. **56** (2004), no. 1-3, 89–113. 13

[BC09] P. J. Bickel and A. Chen, *A nonparametric view of network models and newmangirvan and other modularities*, Proceedings of the National Academy of Sciences (2009). 8

[BCLS87] T.N. Bui, S. Chaudhuri, F.T. Leighton, and M. Sipser, *Graph bisection algorithms with good average case behavior*, Combinatorica **7** (1987), no. 2, 171–191 (English). 8

[BCS15] C. Borgs, J. Chayes, and A. Smith, *Private graphon estimation for sparse graphs*, In preparation (2015). 13

[Bes75] J. Besag, *Statistical analysis of non-lattice data*, The Statistician **24** (1975), no. 3, 179–195. 2

[Bes77] ———, *Efficiency of pseudolikelihood estimation for simple Gaussian fields*, Biometrika **64** (1977), no. 3, 616–618. 2

[BH14] J. Xu B. Hajek, Y. Wu, *Achieving exact cluster recovery threshold via semidefinite programming*, arXiv:1412.6156 (2014). 13

[BHT89] A. Buja, T. J. Hastie, and R. Tibshirani, *Linear smoothers and additive models*, Annals of Statistics **17** (1989), no. 2, 453–510. 5

[Bir83] L. Birgé, *Approximation dans les espaces metriques et theorie de l'estimation*, Z. Wahrsch. verw. Gebiete **65** (1983), 181–327. 4

[Bir87] ———, *Estimating a density under order restrictions: Non-asymptotic minimax risk*, Annals of Statistics **15** (1987), no. 3, 995–1012. 4

[BLM13] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Oxford University Press, Oxford, UK, 2013. 1

[BM09] J. Bento and A. Montanari, *Which graphical models are difficulty to learn?*, Proceedings of the NIPS Conference, December 2009. 2

[BMS13] G. Bresler, E. Mossel, and A. Sly, *Reconstruction of Markov Random Fields from samples: Some observations and algorithms*, SIAM Journal on Computing **42** (2013), no. 2, 563–578. 1, 2

[Bop87] R.B. Boppana, *Eigenvalues and graph bisection: An average-case analysis*, In 28th Annual Symposium on Foundations of Computer Science (1987), 280–285. 8

[BR13] Q. Berthet and P. Rigollet, *Computational lower bounds for sparse PCA*, Conference on Computational Learning Theory (Princeton, NJ), June 2013. 3

[Bre14] G. Bresler, *Efficiently learning Ising models on arbitrary graphs*, Tech. report, MIT, 2014. 2

[BRT09] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, Annals of Statistics **37** (2009), no. 4, 1705–1732. 5

[BS04] B. Bollobás and A. D. Scott, *Max cut for random graphs with a planted partition*, Comb. Probab. Comput. **13** (2004), no. 4-5, 451–474. 8

[CG14] Y. Chen and A. J. Goldsmith, *Information recovery from pairwise measurements*, In Proc. ISIT, Honolulu. (2014). 13

[CHG] Y. Chen, Q.-X. Huang, and L. Guibas, *Near-optimal joint object matching via convex relaxation*, Available Online: arXiv:1402.1473 [cs.LG]. 13

[CI01] T. Carson and R. Impagliazzo, *Hill-climbing finds random planted bisections*, Proc. 12th Symposium on Discrete Algorithms (SODA 01), ACM press, 2001, 2001, pp. 903–909. 8

[CK99] A. Condon and R. M. Karp, *Algorithms for graph partitioning on the planted partition model*, Lecture Notes in Computer Science **1671** (1999), 221–232. 8

[CL68] C. K. Chow and C. N. Liu, *Approximating discrete probability distributions with dependence trees*, IEEE Trans. Info. Theory **IT-14** (1968), 462–467. 1

[Co10] A. Coja-oghlan, *Graph partitioning via adaptive spectral techniques*, Comb. Probab. Comput. **19** (2010), no. 2, 227–284. 8, 9

[CRV15] P. Chin, A. Rao, and V. Vu, *Stochastic block model and community detection in the sparse graphs: A spectral algorithm with optimal rate of recovery*, arXiv:1501.05021 (2015). 13

[CSX12] Y. Chen, S. Sanghavi, and H. Xu, *Clustering Sparse Graphs*, arXiv:1210.3335 (2012). 12

[CT91] T.M. Cover and J.A. Thomas, *Elements of information theory*, John Wiley and Sons, New York, 1991. 1

[CT05] E. J. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. Info Theory **51** (2005), no. 12, 4203–4215. 5

[CT06] I. Csiszár and Z. Talata, *Consistent estimation of the basic neighborhood structure of Markov random fields*, The Annals of Statistics **34** (2006), no. 1, 123–145. 2

[CT07] E. J. Candès and T. Tao, *The Dantzig selector: Statistical estimation when p is much larger than n*, Annals of Statistics **35** (2007), no. 6, 2313–2351. 5

[CWA12] D. S. Choi, P. J. Wolfe, and E. M. Airoldi, *Stochastic blockmodels with a growing number of classes*, Biometrika (2012). 8, 13

[CY06] J. Chen and B. Yuan, *Detecting functional modules in the yeast proteinprotein interaction network*, Bioinformatics **22** (2006), no. 18, 2283–2290. 8

[DAM15] Y. Deshpande, E. Abbe, and A. Montanari, *Asymptotic mutual information for the two-groups stochastic block model*, In preparation (2015). 12, 14, 15

[dEJL07] A. d'Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet, *A direct formulation for sparse PCA using semidefinite programming*, SIAM Review **49** (2007), no. 3, 434–448. 3

[DF89] M.E. Dyer and A.M. Frieze, *The solution of some random NP-hard problems in polynomial expected time*, Journal of Algorithms **10** (1989), no. 4, 451 – 489. 8

[DKMZ11] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications*, Phys. Rev. E **84** (2011), 066106. 8, 9, 11, 12, 15

[DL91] D. L. Donoho and R. Liu, *Geometrizing rates of convergence II*, Annals of Statistics **19** (1991), 633–667. 4

[DL93] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Springer-Verlag, New York, NY, 1993. 4

[DM13] Y. Deshpande and A. Montanari, *Finding hidden cliques of size n/e in nearly linear time*, arXiv:1304.7047 (2013). 13

[DM14] Yash Deshpande and Andrea Montanari, *Information-theoretically optimal sparse pca*, Information Theory (ISIT), 2014 IEEE International Symposium on, IEEE, 2014, pp. 2197–2201. 14, 15

[Don06a] D. L. Donoho, *For most large underdetermined systems of linear equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution*, Communications on Pure and Applied Mathematics **59** (2006), no. 7, 907–934. 5

[Don06b] ———, *For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution*, Communications on Pure and Applied Mathematics **59** (2006), no. 6, 797–829. 5

[EKPS00] W. Evans, C. Kenyon, Y. Peres, and L. J. Schulman, *Broadcasting on trees and the Ising model*, Ann. Appl. Probab. **10** (2000), 410–433. 12

[EL07] P. P. B. Eggermont and V. N. LaRiccia, *Maximum penalized likelihood estimation: V. ii regression*, Springer Series in Statistics, vol. 2, Springer, New York, NY, 2007. 4

[FMW85] S. E. Fienberg, M. M. Meyer, and S. S. Wasserman, *Statistical analysis of multiple sociometric relations*, Journal of The American Statistical Association (1985), 51–67. 8

[For10] S. Fortunato, *Community detection in graphs*, Physics Reports **486 (3-5)** (2010), 75–174. 6

[FRG09] A. K. Fletcher, S. Rangan, and V. K. Goyal, *Necessary and sufficient conditions for sparsity pattern recovery*, IEEE Transactions on Information Theory **55** (2009), no. 12, 5758–5772. 2

[GB13] P. K. Gopalan and D. M. Blei, *Efficient discovery of overlapping communities in massive networks*, Proceedings of the National Academy of Sciences (2013). 9

[GKKW02] L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk, *A distribution-free theory of nonparametric regression*, Springer Series in Statistics, Springer, 2002. 4

[GN02] M. Girvan and M. E. J. Newman, *Community structure in social and biological networks*, Proceedings of the National Academy of Sciences **99** (2002), no. 12, 7821–7826. 8

[GR04] E. Greenshtein and Y. Ritov, *Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization*, Bernoulli **10** (2004), 971–988. 5

[GRSY14] A. Globerson, T. Roughgarden, D. Sontag, and C. Yildirim, *Tight error bounds for structured prediction*, CoRR **abs/1409.5834** (2014). 13

[GSV05] Dongning Guo, Shlomo Shamai, and Sergio Verdú, *Mutual information and minimum mean-square error in gaussian channels*, Information Theory, IEEE Transactions on **51** (2005), no. 4, 1261–1282. 15

[Gun11] A. Guntuboyina, *Lower bounds for the minimax risk using f-divergences and applications*, IEEE Transactions on Information Theory **57** (2011), no. 4, 2386–2399. 4

[GV14] O. Guédon and R. Vershynin, *Community detection in sparse networks via Grothendieck's inequality*, ArXiv:1411.4686 (2014). 13

[Has78] R. Z. Hasminskii, *A lower bound on the risks of nonparametric estimates of densities in the uniform metric*, Theory Prob. Appl. **23** (1978), 794–798. 4

[HG13] Q.-X. Huang and L. Guibas, *Consistent shape maps via semidefinite programming*, Computer Graphics Forum **32** (2013), no. 5, 177–186. 13

[HI90] R. Z. Hasminskii and I. Ibragimov, *On density estimation in the view of Kolmogorov's ideas in approximation theory*, Annals of Statistics **18** (1990), no. 3, 999–1010. 4

[HLL83] P. W. Holland, K. Laskey, and S. Leinhardt, *Stochastic blockmodels: First steps*, Social Networks **5** (1983), no. 2, 109–137. 8

[HLM12] S. Heimlicher, M. Lelarge, and L. Massoulié, *Community detection in the labelled stochastic block model*, arXiv:1209.2910 (2012). 13

[HMSW04] W. K. Härdle, M. Müller, S. Sperlich, and A. Werwatz, *Nonparametric and semiparametric models*, Springer Series in Statistics, Springer, New York, 2004. 4

[HT86] T. Hastie and R. Tibshirani, *Generalized additive models*, Statistical Science **1** (1986), no. 3, 297–310. 5

[Isi25] E. Ising, *Beitrag zur theorie der ferromagnetismus*, Zeitschrift für Physik **31** (1925), no. 1, 253–258. 1

[JL09a] I. M. Johnstone and A. Y. Lu, *On consistency and sparsity for principal components analysis in high dimensions*, Journal of the American Statistical Association **104** (2009), 682–693. 3

[JL09b] Iain M Johnstone and Arthur Yu Lu, *On consistency and sparsity for principal components analysis in high dimensions*, Journal of the American Statistical Association **104** (2009), no. 486. 14

[Joh01] I. M. Johnstone, *On the distribution of the largest eigenvalue in principal components analysis*, Annals of Statistics **29** (2001), no. 2, 295–327. 3

[Jol04] I. T. Jolliffe, *Principal Component Analysis*, Springer, New York, NY, 2004. 3

[JS98] Mark Jerrum and Gregory B. Sorkin, *The metropolis algorithm for graph bisection*, Discrete Applied Mathematics **82** (1998), no. 13, 155 – 175. 8

[JTZ04] D. Jiang, C. Tang, and A. Zhang, *Cluster analysis for gene expression data: a survey*, Knowledge and Data Engineering, IEEE Transactions on **16** (2004), no. 11, 1370–1386. 8

[KB07] M. Kalisch and P. Bühlmann, *Estimating high-dimensional directed acyclic graphs with the PC algorithm*, Journal of Machine Learning Research **8** (2007), 613–636. 1

[KF10] D. Koller and N. Friedman, *Graphical models*, MIT Press, New York, 2010. 1

[KN11] B. Karrer and M. E. J. Newman, *Stochastic blockmodels and community structure in networks*, Phys. Rev. E **83** (2011), 016107. 8, 13

[KPSS10] K. R. Kumar, P. Pakzad, A.H. Salavati, and A. Shokrollahi, *Phase transitions for mutual information*, Turbo Codes and Iterative Information Processing (ISTC), 2010 6th International Symposium on, 2010, pp. 137–141. 13

[KS01] D. Karger and N. Srebro, *Learning Markov networks: maximum bounded tree-width graphs*, Symposium on Discrete Algorithms, 2001, pp. 392–401. 1

[KT59] A. Kolmogorov and B. Tikhomirov, *ϵ-entropy and ϵ-capacity of sets in functional spaces*, Uspekhi Mat. Nauk. **86** (1959), 3–86, Appeared in English as Amer. Math. Soc. Translations, 17:277–364, 1961. 1, 4, 5

[KY10] V. Koltchinskii and M. Yuan, *Sparsity in multiple kernel learning*, Annals of Statistics **38** (2010), 3660–3695. 5

[LC73] L. Le Cam, *Convergence of estimates under dimensionality restrictions*, Annals of Statistics (1973). 4

[Led01] M. Ledoux, *The Concentration of Measure Phenomenon*, Mathematical Surveys and Monographs, American Mathematical Society, Providence, RI, 2001. 1

[Lov12] L. Lovász, *Large networks and graph limits*, American Mathematical Society colloquium publications, American Mathematical Society, 2012. 13

[LSY03] G. Linden, B. Smith, and J. York, *Amazon.com recommendations: Item-to-item collaborative filtering*, IEEE Internet Computing **7** (2003), no. 1, 76–80. 8

[Mas14] L. Massoulié, *Community detection thresholds and the weak Ramanujan property*, STOC 2014: 46th Annual Symposium on the Theory of Computing (New York, United States), June 2014, pp. 1–10. 8, 9, 15

[MB06] N. Meinshausen and P. Bühlmann, *High-dimensional graphs and variable selection with the Lasso*, Annals of Statistics **34** (2006), 1436–1462. 2

[McS01] F. McSherry, *Spectral partitioning of random graphs*, In 42nd Annual Symposium on Foundations of Computer Science (2001), 529–537. 8

[MNSa] E. Mossel, J. Neeman, and A. Sly, *Belief propagation, robust reconstruction, and optimal recovery of block models*, Arxiv:arXiv:1309.1380. 9, 12

[MNSb] ———, *Consistency thresholds for binary symmetric block models*, Arxiv:arXiv:1407.1591. To appear in STOC15. 8, 9, 15

[MNS12] E. Mossel, J. Neeman, and A. Sly, *Stochastic block models and reconstruction*, Available online at arXiv:1202.1499 [math.PR] (2012). 12, 15

[MNS14] ———, *A proof of the block model threshold conjecture*, Available online at arXiv:1311.4115 [math.PR] (2014). 8, 9, 15

[Mon15] A. Montanari, *Finding one community in a sparse graph*, arXiv:1502.05680 (2015). 13

[MPN+99] E.M. Marcotte, M. Pellegrini, H.-L. Ng, D.W. Rice, T.O. Yeates, and D. Eisenberg, *Detecting protein function and protein-protein interactions from genome sequences*, Science **285** (1999), no. 5428, 751–753. 8

[MvdGB09] L. Meier, S. van de Geer, and P. Bühlmann, *High-dimensional additive modeling*, Annals of Statistics **37** (2009), 3779–3821. 5

[MW13] Z. Ma and Y. Wu, *Computational barriers in minimax submatrix detection*, arXiv preprint arXiv:1309.5914 (2013). 3

[NBSS10] P. Netrapalli, S. Banerjee, S. Sanghavi, and S. Shakkottai, *Greedy learning of Markov network structure*, Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on, IEEE, 2010, pp. 1295–1302. 2

[Nem00] A. Nemirovski, *Topics in non-parametric statistics*, Ecole d'été de Probabilities de Saint-Flour XXVIII (P. Bernard, ed.), Lecture notes in Mathematics, Springer, 2000. 5

[New10] M. Newman, *Networks: an introduction*, Oxford University Press, Oxford, 2010.

[NRWY12] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, *A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers*, Statistical Science **27** (2012), no. 4, 538–557. 1, 5

[NWS] M. E. J. Newman, D. J. Watts, and S. H. Strogatz, *Random graph models of social networks*, Proc. Natl. Acad. Sci. USA **99**, 2566–2572. 8

[Pin85] A. Pinkus, *N-widths in approximation theory*, Springer, New York, 1985. 4

[PM14] G. Puleo and O. Milenkovic, *Correlation clustering with constrained cluster sizes and extended weights bounds*, arXiv:1411.0547 (2014). 13

[PSD00] J. K. Pritchard, M. Stephens, and P. Donnelly, *Inference of Population Structure Using Multilocus Genotype Data*, Genetics **155** (2000), no. 2, 945–959. 8

[RCY11] K. Rohe, S. Chatterjee, and B. Yu, *Spectral clustering and the high-dimensional stochastic blockmodel*, The Annals of Statistics **39** (2011), no. 4, 1878–1915. 8

[RDRI+14] C. Rudin, D. Dunson, H. Ji R. Irizarry, E. Laber, J. Leek, T. McCormick, S. Rose, C. Schafer, M. van der Laan, L. Wasserman, and L. Xue, *Discovery with data: Leveraging statistics with computer science to transform science and society*, American Statistical Association (2014). 8

[RLLW09] P. Ravikumar, H. Liu, J. D. Lafferty, and L. A. Wasserman, *SpAM: sparse additive models*, Journal of the Royal Statistical Society, Series B **71** (2009), no. 5, 1009–1030. 5

[RWL10] P. Ravikumar, M. J. Wainwright, and J. D. Lafferty, *High-dimensional Ising model selection using ℓ_1-regularized logistic regression*, Annals of Statistics **38** (2010), no. 3, 1287–1319. 2

[RWY11] G. Raskutti, M. J. Wainwright, and B. Yu, *Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls*, IEEE Trans. Information Theory **57** (2011), no. 10, 6976—6994. 5

[RWY12] ———, *Minimax-optimal rates for sparse additive models over kernel classes via convex programming*, Journal of Machine Learning Research **12** (2012), 389–427. 5

[RWY14] ———, *Early stopping and non-parametric regression: An optimal data-dependent stopping rule*, Journal of Machine Learning Research **15** (2014), 335–366. 5

[Sch78] G. Schwarz, *Estimating the dimension of a model*, Annals of Statistics **6** (1978), 461–468. 2

[SGS00] P. Spirtes, C. Glymour, and R. Scheines, *Causation, prediction and search*, MIT Press (2000). 1

[SHB07] M. Sonka, V. Hlavac, and R. Boyle, *Image processing, analysis, and machine vision*, Thomson-Engineering, 2007. 8

[SKLZ15] A. Saade, F. Krzakala, M. Lelarge, and L. Zdeborová, *Spectral detection in the censored block model*, arXiv:1502.00163 (2015). 13

[SM97] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (1997), 888–905. 6, 8

[SN97] T. A. B. Snijders and K. Nowicki, *Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure*, Journal of Classification **14** (1997), no. 1, 75–100. 8

[SPT+01] T. Sorlie, C.M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, Mi.B. Eisen, M. van de Rijn, S.S. Jeffrey, T. Thorsen, H. Quist, J.C. Matese, P.O. Brown, D. Botstein, P.E. Lonning, and A. Borresen-Dale, *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications*, no. 19, 10869–10874. 8

[Sto85] C. J. Stone, *Additive regression and other non-parametric models*, Annals of Statistics **13** (1985), no. 2, 689–705. 5

[Tsy09] A. B. Tsybakov, *Introduction to nonparametric estimation*, Springer, New York, 2009. 4, 5

[vdG00] S. van de Geer, *Empirical processes in m-estimation*, Cambridge University Press, 2000. 4

[vdGB09] S. van de Geer and P. Bühlmann, *On the conditions used to prove oracle results for the Lasso*, Electronic Journal of Statistics **3** (2009), 1360–1392. 5

[Wai09a] M. J. Wainwright, *Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting*, IEEE Trans. Info. Theory **55** (2009), 5728–5741. 2

[Wai09b] ———, *Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)*, IEEE Trans. Information Theory **55** (2009), 2183–2202. 2

[Wai14a] ———, *Constrained forms of statistical minimax: Computation, communication and privacy*, Proceedings of the International Congress of Mathematicians (Seoul, Korea), 2014. 4

[Wai14b] ———, *Structured regularizers: Statistical and computational issues*, Annual Review of Statistics and its Applications **1** (2014), 233–253. 1

[Wai15] ———, *High-dimensional statistics: A non-asymptotic viewpoint*, Cambridge University Press, 2015. 4

[Was06] L. A. Wasserman, *All of Non-Parametric Statistics*, Springer Series in Statistics, Springer-Verlag, New York, NY, 2006. 4, 5

[WBB76] H. C. White, S. A. Boorman, and R. L. Breiger, *Social structure from multiple networks*, American Journal of Sociology **81** (1976), 730–780. 8

[WJ08] M. J. Wainwright and M. I. Jordan, *Graphical models, exponential families and variational inference*, Foundations and Trends in Machine Learning **1** (2008), no. 1–2, 1—305. 1

[WW87] Y. J. Wang and G. Y. Wong, *Stochastic blockmodels for directed graphs*, Journal of the American Statistical Association (1987), 8–19. 8

[XLM14] J. Xu, M. Lelarge, and L. Massoulie, *Edge label inference in generalized stochastic block models: from spectral theory to impossibility results*, to appear in Proceedings of COLT 2014 (2014). 13

[XWZ+14] J. Xu, R. Wu, K. Zhu, B. Hajek, R. Srikant, and L. Ying, *Jointly clustering rows and columns of binary matrices: Algorithms and trade-offs*, SIGMETRICS Perform. Eval. Rev. **42** (2014), no. 1, 29–41. 8

[YB99] Y. Yang and A. Barron, *Information-theoretic determination of minimax rates of convergence*, Annals of Statistics **27** (1999), no. 5, 1564–1599. 4

[YC14] J. Xu Y. Chen, *Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices*, arXiv:1402.1267 (2014). 8, 9, 12, 13

[YP14] S. Yun and A. Proutiere, *Accurate community detection in the stochastic block model via spectral algorithms*, arXiv:1412.7335 (2014). 13

[Yu93] B. Yu, *Assouad, Fano and Le Cam*, Festschrift in Honor of L. Le Cam on his 70th Birthday, 1993. 4

[ZWJ14] Y. Zhang, M. J. Wainwright, and M. I. Jordan, *Lower bounds on the performance of polynomial-time algorithms for sparse linear regression*, Proceedings of the Conference on Learning Theory (COLT) (Barcelona, Spain), June 2014, Full length version at http://arxiv.org/abs/1402.1918. 5

[ZWJ15] ———, *Optimal prediction for sparse linear models? Lower bounds for coordinate-separable M-estimators*, Tech. report, UC Berkeley, March 2015, arxiv:1503.03188. 5

[ZY06] P. Zhao and B. Yu, *On model selection consistency of Lasso*, Journal of Machine Learning Research **7** (2006), 2541–2567. 2