Report from Dagstuhl Seminar 18112

Coding Theory for Inference, Learning and Optimization

Edited by

Po-Ling Loh[1], Arya Mazumdar[2], Dimitris Papailiopoulos[3], and Rüdiger Urbanke[4]

1 **University of Wisconsin – Madison, US,** loh@ece.wisc.edu
2 **University of Massachusetts, US,** arya@cs.umass.edu
3 **University of Wisconsin – Madison, US,** dimitris@papail.io
4 **EPFL – Lausanne, CH,** rudiger.urbanke@epfl.ch

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 18112, "Coding Theory for Inference, Learning and Optimization."

Coding theory has recently found new applications in areas such as distributed machine learning, dimension reduction, and variety of statistical problems involving estimation and inference. In machine learning applications that use large-scale data, it is desirable to communicate the results of distributed computations in an efficient and robust manner. In dimension reduction applications, the pseudorandom properties of algebraic codes may be used to construct projection matrices that are deterministic and facilitate algorithmic efficiency. Finally, relationships that have been forged between coding theory and problems in theoretical computer science, such as k-SAT or the planted clique problem, lead to a new interpretation of the sharp thresholds encountered in these settings in terms of thresholds in channel coding theory.

The aim of this Dagstuhl Seminar was to draw together researchers from industry and academia that are working in coding theory, particularly in these different (and somewhat disparate) application areas of machine learning and inference. The discussions and collaborations facilitated by this seminar were intended to spark new ideas about how coding theory may be used to improve and inform modern techniques for data analytics.

Seminar March 11–16, 2018 – https://www.dagstuhl.de/18112
2012 ACM Subject Classification Mathematics of computing → Coding theory,Mathematics of computing → Probability and statistics, Theory of computation → Mathematical optimization
Keywords and phrases Coding theory, Distributed optimization, Machine learning, Threshold phenomena
Digital Object Identifier 10.4230/DagRep.8.3.60
Edited in cooperation with Po-Ling Loh

1 Executive Summary

Po-Ling Loh (University of Wisconsin – Madison, US)
Arya Mazumdar (University of Massachusetts, US)
Dimitris Papailiopoulos (University of Wisconsin – Madison, US)
Rüdiger Urbanke (EPFL – Lausanne, CH)

Codes are widely used in engineering applications to offer reliability and fault tolerance. The high-level idea of coding is to exploit redundancy in order to create robustness against

system noise. The theoretical properties of codes have been studied for decades both from a purely mathematical point of view, as well as in various engineering contexts. The latter have resulted in constructions that have been incorporated into our daily lives: No storage device, cell phone transmission, or Wi-Fi connection would be possible without well-constructed codes.

Recent research has connected concepts in coding theory to non-traditional applications in learning, computation and inference, where codes have been used to design more efficient inference algorithms and build robust, large-scale, distributed computational pipelines. Moreover, ideas derived from Shannon theory and the algebraic properties of random codes have resulted in novel research that sheds light on fundamental phase transition phenomena in several long-standing combinatorial and graph-theoretic problems.

The main goal of our seminar was to accelerate research in the growing field of coding theory for computation and learning, and maximize the transformative role of codes in non-traditional application areas. The seminar brought together 22 researchers from across the world specializing in information theory, machine learning, theoretical computer science, optimization, and statistics. The schedule for each day included a tutorial talk by a senior researcher, followed by shorter talks by participants on recent or ongoing work. The afternoons were devoted to informal breakout sessions for groups to discuss open questions. Two of the larger breakout sessions focused on distributed optimization and group testing.

Seminar participants reported that they enjoyed hearing about new ideas, as well as delving into deeper technical discussions about open problems in coding theory. Some topics deserving special mention include the use of techniques in statistical mechanics; locally decodable and recoverable codes; submodular function optimization; hypergraph clustering; private information retrieval; and contagion on graphs. All participants valued the ample time for discussions between and after talks, as it provided a fruitful atmosphere for collaborating on new topics.

2 Table of Contents

3 Overview of Talks

3.1 Error-correcting codes as samplers

Dimitris Achlioptas (University of California – Santa Cruz, US)

We consider the task of summing a non-negative function f over a discrete set Ω; e.g., to compute the partition function of a graphical model. Ermon et al. have shown that in a probabilistic, approximate sense, summation can be reduced to maximizing f over random subsets of Ω defined by parity (XOR) constraints. Unfortunately, XORs with many variables are computationally intractable, while XORs with few variables have poor statistical performance. We introduce two ideas to address this tradeoff, both motivated by the theory of error-correcting codes. The first is to maximize f over explicitly generated random affine subspaces of Ω, which is equivalent to unconstrained maximization of f over an exponentially smaller domain. The second idea, closer in spirit to the original approach, is to use systems of linear equations defining Low Density Parity Check (LDPC) codes. Even though the equations in such systems only contain $O(1)$ variables each, their sets of solutions (codewords) have excellent statistical properties. By combining these ideas, we achieve dramatic speedup over the original approach and levels of accuracy that were previously unattainable.

3.2 Different facets of the repair problem

Alexander Barg (University of Maryland – College Park, US)

Joint work of Min Ye, Itzhak Tamo and Alexander Barg

The repair problem refers to correcting one or several erasures with a given error-correcting code using as little inter-nodal communication as possible. An information-theoretic lower bound on the repair bandwidth is known from the literature, and codes that meet it with equality are said to support optimal repair. In this talk, we consider several versions of the repair problem, illustrating different techniques behind the construction of optimal codes. We begin by describing an approach to optimal-repair codes using the notion of interference alignment, and presenting several families of codes with a number of additional properties (universality, error tolerance, optimal updates). We then discuss optimal repair of Reed-Solomon codes, presenting codes from this family that support optimal repair of a single failed node or of several failed nodes. In the final part of the talk, we discuss one more setting of the problem, where the task is to perform repair of several nodes in a distributed way, and again construct a family of codes with optimal repair bandwidth.

3.3 Twisted Reed-Solomon codes: Novel class of MDS codes

Martin Bossert (Universität Ulm, DE)

Joint work of Sven Puchinger, Johan Roseskilde, Peter Beelen, and Martin Bossert

We introduce the class of twisted RS codes and prove the MDS property. We derive bounds for the length of the codes, and show that many twisted RS codes exist which are not equivalent to RS codes (or Roth-Lempel codes). We study the application of twisted RS codes in the McEliece cryptosystem and show that the attacks for RS codes do not work in case of twisted RS codes. Furthermore, we show examples of the number of existing twisted RS codes that are not equivalent to RS codes for several parameters.

3.4 Can we access a database both locally and privately?

Elette Boyle (The Interdisciplinary Center – Herzliya, IL)

A private information retrieval (PIR) protocol allows a client to retrieve an item from a remote database while hiding which item is retrieved even from the servers storing the database. The main focus of the large body of work on PIR has been on minimizing the communication complexity. We present an exploratory approach for achieving PIR with sublinear computational complexity, based on a new primitive of Oblivious Locally Decodable Codes.

3.5 Random linear equations

Amin Coja-Oghlan (Goethe-Universität – Frankfurt am Main, DE)

Joint work of Peter Ayre, Pu Gao, Noela Müller, and Amin Coja-Oghlan

Let A be a random $m \times n$ matrix over the finite field F_q with precisely k non-zero entries per row, and let $y \in F_q^m$ be a random vector chosen independently of A. We identify the threshold m/n up to which the linear system $Ax = y$ has a solution with high probability, and analyze the geometry of the set of solutions. In the special case $q = 2$, known as the random k-XORSAT problem, the threshold was determined by [Dubois and Mandler, 2002; Dietzfelbinger et al., 2010; Pittel and Sorkin, 2016], and the proof technique was subsequently extended to the cases $q = 3, 4$ [Falke and Goerdt, 2012]. But the argument depends on technically demanding second moment calculations that do not generalize to $q > 3$. Here, we approach the problem from the viewpoint of a decoding task, which leads to a transparent combinatorial proof.

3.6 A lower bound for maximally recoverable codes with locality

Venkatesan Guruswami (Carnegie Mellon University – Pittsburgh, US)

Joint work of Sivakanth Gopi, Sergey Yekhanin, and Venkatesan Guruswami

MDS codes like Reed-Solomon codes enable erasure correction with optimal redundancy: with h redundant symbols, they allow recovery of any subset of h erased positions. In matrix terms, they are defined by an $h \times n$ parity check matrix, all of whose $h \times h$ submatrices are nonsingular. Such matrices, e.g., Vandermonde, exist over a field of size $O(n)$.

The prevalent use of erasure coding in today's large distributed storage systems, where individual storage nodes often fail, brings to the fore a new requirement: the ability to quickly recover any single symbol of the codeword based on few other codeword symbols. Such "locality" can be built into the code via local parity checks—for example, the n codeword symbols can be partitioned into n/r groups, each with r symbols, obeying a local parity check. Here, r is the locality parameter of the code. In matrix terms, we have an $(n/r + h) \times n$ matrix with the first n/r rows, each with r 1's, corresponding to the local parity checks, and the last h rows unrestricted. The local checks compromise the MDS property, but we would still like the code to correct all erasure patterns that can possibly be recovered given the specified topology of parity checks. This property is called Maximal Recoverability, and for the above topology, amounts to the nonsingularity of every $(n/r + h) \times (n/r + h)$ submatrix that includes at least one column from each local group.

The known constructions (and even existence proofs) of such matrices require a very large field size of about n^h, and it has been an important question whether MR codes can exist over smaller, even $O(n)$-sized, fields. The talk will mention the prior construction with n^h field size, and then present a recent super-linear lower bound on the field size which relies on known vertex expansion properties of the hyperplane-point incidence graph in projective space.

3.7 Shifted weight distributions

Anna Gál (University of Texas – Austin, US)

We discuss an open problem of a coding-theoretic flavor. This question came up as part of an approach towards solving an open problem in computational complexity theory.

3.8 Submodular maximization: The decentralized setting

Hamed S. Hassani (University of Pennsylvania – Philadelphia, US)

In this talk, we showcase the interplay between discrete and continuous optimization in network-structured settings. We propose the first fully-decentralized optimization method for a wide class of non-convex objective functions that possess a diminishing returns property. More specifically, given an arbitrary connected network and a global continuous submodular

function, formed by a sum of local functions, we develop Decentralized Continuous Greedy (DCG), a message-passing algorithm that converges to the tight $(1-1/e)$-approximation factor of the optimum global solution using only local computation and communication. We also provide strong convergence bounds as a function of network size and spectral characteristics of the underlying topology. Interestingly, DCG readily provides a simple recipe for decentralized discrete submodular maximization through the means of continuous relaxations. Formally, we demonstrate that by lifting the local discrete functions to continuous domains and using DCG as an interface, we can develop a consensus algorithm that also achieves the tight $(1-1/e)$-approximation guarantee of the global discrete solution, once a proper rounding scheme is applied.

3.9 Sufficiently myopic adversaries are weak

Sidharth Jaggi (The Chinese University of Hong Kong, HK)

In this work, we consider a communication problem in which a sender, Alice, wishes to communicate with a receiver, Bob, over a channel controlled by an adversarial jammer, James, who is *myopic*. Roughly speaking, for blocklength n, the codeword X^n transmitted by Alice is corrupted by James, who must base his adversarial decisions (of which locations of X^n to corrupt and how to corrupt them) not on the codeword X^n, but on Z^n, an image of X^n through a noisy memoryless channel. More specifically, our communication model may be described by two channels: A memoryless channel $p(z|x)$ from Alice to James, and an *Arbitrarily Varying Channel* $p(y|x,s)$ from Alice to Bob, governed by a state X^n determined by James. In standard adversarial channels, the states S^n may depend on the codeword X^n, but in our setting, S^n depends only on James's view Z^n.

The myopic channel captures a broad range of channels and bridges between the standard models of memoryless and adversarial (zero-error) channels. In this work, we present upper and lower bounds on the capacity of myopic channels. For a number of special cases of interest, we show that our bounds are tight. We extend our results to the setting of *secure* communication, in which we require that the transmitted message remain secret from James. For example, we show that if (i) James may flip at most a p fraction of the bits communicated between Alice and Bob, and (ii) James views X^n through a binary symmetric channel with parameter q, then once James is "sufficiently myopic" (in this case, when $q > p$), the optimal communication rate is that of an adversary who is "blind" (that is, an adversary that does not see X^n at all), which is $1 - H(p)$ for standard communication, and $H(q) - H(p)$ for secure communication. A similar phenomenon exists for our general model of communication.

3.10 Fundamental limits of symmetric low-rank matrix estimation

Marc Lelarge (ENS – Paris, FR)

Joint work of Leo Miolane and Marc Lelarge

We consider the high-dimensional inference problem, where the signal is a low-rank symmetric matrix which is corrupted by additive Gaussian noise. Given a probabilistic model for the low-rank matrix, we compute the limit in the large-dimension setting for the mutual information between the signal and the observations, as well as the matrix minimum mean square error, while the rank of the signal remains constant. We also show that our model extends beyond the particular case of additive Gaussian noise, and we prove an universality result connecting the community detection problem to our Gaussian framework. We unify and generalize a number of recent works on PCA, sparse PCA, submatrix localization, and community detection, by computing the information-theoretic limits for these problems in the high-noise regime. In addition, we show that the posterior distribution of the signal given the observations is characterized by a parameter of the same dimension as the square of the rank of the signal (i.e., scalar in the case of rank one). Finally, we connect our work with the hard but detectable conjecture in statistical physics.

3.11 Statistical inference for infectious disease modeling

Po-Ling Loh (University of Wisconsin – Madison, US)

Joint work of Justin Khim and Po-Ling Loh

We discuss two recent results concerning disease modeling on networks. The infection is assumed to spread via contagion (e.g., transmission over the edges of an underlying network). In the first scenario, we observe the infection status of individuals at a particular time instance and the goal is to identify a confidence set of nodes that contain the source of the infection with high probability. We show that when the underlying graph is a tree with certain regularity properties and the structure of the graph is known, confidence sets may be constructed with cardinality independent of the size of the infection set. In the second scenario, the goal is to infer the network structure of the underlying graph based on knowledge of the infected individuals. We develop a hypothesis test based on permutation testing, and describe a sufficient condition for the validity of the hypothesis test based on automorphism groups of the graphs involved in the hypothesis test.

3.12 The adaptive interpolation method for proving replica formulas

Nicolas Macris (EPFL – Lausanne, CH)

In this talk, we give an introduction to the newly-introduced adaptive interpolation method to prove in a simple and unified way replica formulas for Bayesian optimal inference problems.

We illustrate the method with a paradigmatic inference problem, namely rank-one matrix estimation.

The replica method from statistical mechanics has been applied to Bayesian inference problems (e.g., coding, estimation) already two decades ago. Rigorous proofs of the formulas for the mutual informations/entropies/free energies stemming from this method have for a long time only been partial, consisting generally of one-sided bounds. It is only quite recently that there has been a surge of progress using various methods (e.g., spatial coupling, the Aizenman-Sims-Starr principle, and the cavity method) to derive full proofs, but which are typically quite complicated. Recently with Jean Barbier, we introduced a powerful evolution of the Guerra-Toninelli interpolation method—called adaptive interpolation—that allows to fully prove the replica formulas in a quite simple and unified way for Bayesian inference problems (we note that in its original, more complicated incarnation, we called this method "stochastic interpolation").

We review the method for the rank-one matrix estimation or factorisation problem (one of the simplest non-linear estimation problems). The main new ingredient is the concentration of the "overlap" which, remarkably, can be proven for Bayesian inference problems when the prior and hyperparameters are all known. We will refer to this setting where the prior and hyperparameters are known as the Bayesian optimal inference.

The adaptive interpolation method has been fruitfully applied to a range of more difficult problems with a dense underlying graphical structure. So far, these include matrix and tensor factorisation, estimation of traditional and generalised linear models, and learning (e.g., compressed sensing and the single-layer perceptron network). For inference problems with an underlying sparse graphical structure, full proofs of replica formulas are scarce and much more involved. The adaptive interpolation method is still in its infancy for sparse systems, and it would be desirable to develop it further.

3.13 Interactive learning for clustering and community detection

Arya Mazumdar (University of Massachusetts, US)

Clustering has always been a central problem of multiple disciplines within computer science, optimization, and statistics. In the model of interactive clustering, a clustering algorithm can adaptively query a (possibly noisy) oracle, with a small number of elements from the set that is to be clustered. For example, the oracle may answer pairwise queries of the form, "do two elements u and v belong to the same cluster?" This model fits exactly to the recently popular experimental setups where crowdsourcing is used to improve clustering accuracy for various data mining tasks. The goal is to minimize the number of such queries while obtaining the optimum clustering. One of our main contributions is to show the power of side information in the form of a similarity matrix: a matrix that represents noisy pairwise relationships, such as one computed by some automated function on attributes of the elements. The reduction in query complexity under general models of the similarity matrix is stunning, and this remains true even when the answer of each query can be erroneous with a certain probability and "resampling" is not allowed. Our results include a general framework for proving lower bounds for classification problems in the interactive setting. Our algorithms are computationally efficient and are parameter-free; i.e., it works without any knowledge of

the number of clusters or the similarity matrix distribution. The query models we propose have interesting connections to popular community detection models such as the stochastic block model.

3.14 Query and higher-order clustering: Some open problems

Olgica Milenkovic (University of Illinois – Urbana Champaign, US)

We will describe a number of problems in clustering and hypergraph clustering that combine discrete optimization and combinatorial techniques to solve learning problems with side information. In particular, we will discuss the connections between the query k-means clustering problem and generalized constrained coupon collector problems, and submodular hypergraph partitioning.

3.15 Representation learning and signal recovery in nonlinear models

Ankit Singh Rawat (MIT – Cambridge, US)

Nonlinear generative models have become increasingly important in capturing the datasets that appear in various domains and realizing various inference tasks around these datasets. In this talk, we present results towards realizing two basic tasks of representation learning and robust signal recovery in the context of nonlinear generative models. In particular, we discuss the estimation of a low-rank matrix from its nonlinear transformation and recovery of a latent signal from its nonlinear measurements in the presence of outliers. The recovery algorithm is agnostic to underlying nonlinearity and comes with a tight performance analysis.

3.16 Coded gradient computation from cyclic MDS codes and expander graphs

Itzhak Tamo (Tel Aviv University, IL)

We discuss cyclic MDS codes and expander graph techniques for distributed gradient computation in the presence of stragglers. For exact gradient computation, the suggested techniques employ cyclic MDS codes to attain comparable parameters to previously known ones, but in a deterministic fashion. For estimated gradient computation, a novel scheme is suggested, stemming from adjacency matrices of expander graphs.

3.17 Inference, coding, and learning in quantum information processing

Pascal Vontobel (The Chinese University of Hong Kong, HK)

Some of the most interesting quantities associated with a factor graph are its marginals and its partition sum. For factor graphs *without cycles* and moderate message-update complexities, the sum-product algorithm (SPA) can be used to efficiently compute these quantities exactly. Moreover, for various classes of factor graphs *with cycles*, the SPA has been successfully applied to efficiently compute good approximations to these quantities. Note that in the case of factor graphs with cycles, the local functions are usually non-negative real-valued functions.

In this talk, we introduce a class of factor graphs, called double-edge factor graphs (DE-FGs), which allow local functions to be complex-valued and only require them, in some suitable sense, to be positive semi-definite kernel functions. We discuss various properties of the SPA when running it on DE-FGs and we show promising numerical results for various example DE-FGs, some of which have connections to quantum information processing.

3.18 Algorithmic applications of list-recovery

Mary Wootters (Stanford University, US)

List-recoverable codes and algorithms for list-recovery have found many applications in algorithm design. In this talk, we define list-recovery, mention state-of-the-art methods for list-recoverable codes, and give two examples of algorithmic applications in group testing and streaming algorithms. The hope is that list-recovery may be helpful in this workshop, as we think about algorithmic applications in inference, optimization, and learning.

4 Working groups

4.1 Group testing

Arya Mazumdar (University of Massachusetts, US)

The Tuesday afternoon break-out session focused on group testing: the recovery of a sparse binary signal under Boolean measurements. The main objective in nonadaptive group testing is to identify a set of defective elements by 1) creating pools of elements; and 2) testing the pools simultaneously for the existence of any defective elements in the pool.

The discussion was led by Sidharth Jaggi, who presented a nice overview of existing signal recovery algorithms for group testing. Earlier in the day, Mary Wootters introduced the group testing problem in relation to list recovery of error-correcting codes. Sidharth continued this discussion by presenting the simplest possible group testing algorithm, which

he called combinatorial orthogonal matching pursuit, or COMP (to acknowledge the well-known orthogonal matching pursuit algorithm from the compressed sensing literature). In the COMP algorithm, all the elements in the pools that produce negative test results are cleared, and remaining items are declared to be defective. Sidharth showed the performance of random test matrices under COMP, and proposed a new heuristic algorithm.

The new algorithm is a generalization of COMP, where the correlations of pairs of columns with the test results are computed (instead of being computed column-wise, as in COMP). This algorithm demonstrates remarkable improvements in empirical simulations; however, the theoretical justification is still far from complete. Sidharth presented several possible ideas to analyze the algorithm, including technical loopholes that would need to be closed.

4.2 Large-scale machine learning meets coding theory

Dimitris Papailiopoulos (University of Wisconsin – Madison, US)

During the Monday afternoon breakout session, we discussed problems related to distributed machine learning and coding theory, focusing on straggler nodes and communication bottlenecks in the context of distributed gradient-based algorithms.

In synchronous distributed setups, the presence of straggler nodes (i.e., nodes slower than the average) can significantly impair the performance of training algorithms. A large body of recent work has focused on reducing the effect of stragglers. Current approaches include replicating jobs across nodes and dropping stragglers in settings where the system can tolerate errors. More recently, coding theory has gained traction as a way to speed up distributed computation. Codes have been used to reduce the runtime of the shuffling phase of MapReduce, improve the efficiency of distributed matrix multiplication, and reduce the effect of stragglers during gradient-based learning. Some interesting problems lying in the intersection between distributed learning and coding theory revolve around statistical accuracy, redundancy, and resilience to stragglers. We discussed several of these problems during our Monday break-out session.

We also discussed communication bottlenecks arising during distributed learning. In gradient-based algorithms, where a master node stores the global model and compute nodes evaluate gradients, frequent communication between nodes is needed to train an accurate model. However, such frequent communication may lead to bottlenecks in the system.

An open problem in the area of communication-efficient distributed training is that of gradient quantization. The challenge of optimally quantizing gradients may be posed as an optimization problem, where the objective is to minimize the variance of stochastic gradients, which controls the rate of convergence, subject to the constraint that gradients are sufficiently simple to represent. The constraints of this optimization problem usually take the following form: 1) the quantized gradient has to be an unbiased estimate of the true gradient on the data; and 2) such quantized gradients need to be represented using few atoms (e.g., bits, elements, or low-rank components). The problem of devising an optimal quantization scheme that minimizes gradient variance, while ensuring an unbiased and compressed representation gradient estimates, remains open. As we discussed in the break-out session, a potential approach to this problem is to establish a connection with traditional element and vector quantization schemes on Gaussian sources.

Participants

- Dimitris Achlioptas
University of California – Santa Cruz, US
- Alexander Barg
University of Maryland – College Park, US
- Martin Bossert
Universität Ulm, DE
- Elette Boyle
The Interdisciplinary Center – Herzliya, IL
- Amin Coja-Oghlan
Goethe-Universität – Frankfurt am Main, DE
- Anna Gál
University of Texas – Austin, US
- Venkatesan Guruswami
Carnegie Mellon University – Pittsburgh, US
- Hamed S. Hassani
University of Pennsylvania – Philadelphia, US
- Sihuang Hu
RWTH Aachen, DE
- Sidharth Jaggi
The Chinese University of Hong Kong, HK
- Marc Lelarge
ENS – Paris, FR
- Po-Ling Loh
University of Wisconsin – Madison, US
- Nicolas Macris
EPFL – Lausanne, CH
- Arya Mazumdar
University of Massachusetts – Amherst, US
- Olgica Milenkovic
University of Illinois – Urbana Champaign, US
- Dimitris Papailiopoulos
University of Wisconsin – Madison, US
- Ankit Singh Rawat
MIT – Cambridge, US
- Changho Suh
KAIST – Daejeon, KR
- Itzhak Tamo
Tel Aviv University, IL
- Rüdiger Urbanke
EPFL – Lausanne, CH
- Pascal Vontobel
The Chinese University of Hong Kong, HK
- Mary Wootters
Stanford University, US