Basic models and questions in statistical network analysis

Miklós Z. Rácz ∗
with Sébastien Bubeck †

Draft of June 7, 2016

Abstract

Extracting information from large graphs has become an important statistical problem since network data is now common in various fields. In this minicourse we will investigate the most natural statistical questions for three canonical probabilistic models of networks: (i) community detection in the stochastic block model, (ii) finding the embedding of a random geometric graph, and (iii) finding the original vertex in a preferential attachment tree. Along the way we will cover many interesting topics in probability theory such as Pólya urns, large deviation theory, concentration of measure in high dimension, entropic central limit theorems, and more.

Outline:

• Lecture 1: A primer on exact recovery in the general stochastic block model.
• Lecture 2: Estimating the dimension of a random geometric graph on a high-dimensional sphere.
• Lecture 3: Introduction to entropic central limit theorems and a proof of the fundamental limits of dimension estimation in random geometric graphs.
• Lectures 4 & 5: Confidence sets for the root in uniform and preferential attachment trees.

∗Microsoft Research; miracz@microsoft.com.
†Microsoft Research; sebubeck@microsoft.com.
1 Lecture 1: A primer on exact recovery in the general stochastic block model

Community detection is a fundamental problem in many sciences, such as sociology (e.g., finding tight-knit groups in social networks), biology (e.g., detecting protein complexes), and beyond. Given its importance, there have been a plethora of algorithms developed in the past few decades to detect communities. But how can we test whether an algorithm performs well? What are the fundamental limits to any community detection algorithm? Often in real data the ground truth is not known (or there is not even a well-defined ground truth), so judging the performance of algorithms can be difficult. Probabilistic generative models can be used to model real networks, and even if they do not fit the data perfectly, they can still be useful: they can act as benchmarks for comparing different clustering algorithms, since the ground truth is known.

Perhaps the most widely studied generative model that exhibits community structure is the stochastic block model (SBM). The SBM was first introduced in sociology [13] and was then studied in several different scientific communities, including mathematics, computer science, physics, and statistics [6, 7, 11, 15, 21]. It gives a distribution on graphs with \( n \) vertices with a hidden partition of the nodes into \( k \) communities. The relative sizes of the communities, and the edge densities connecting communities are parameters of the general SBM. The statistical inference problem is then to recover as much of the community structure as possible given a realization of the graph, but without knowing any of the community labels.

1.1 The stochastic block model and notions of recovery

The general stochastic block model is a distribution on graphs with latent community structure, and it has three parameters: \( n \), the number of vertices; a probability distribution \( p = (p_1, \ldots, p_k) \) that describes the relative sizes of the communities; and \( Q \in [0,1]^{k \times k} \), a symmetric \( k \times k \) matrix that describes the probabilities with which two given vertices are connected, depending on which communities they belong to. The number of communities, \( k \), is implicit in this notation; in these notes we assume that \( k \) is a fixed constant. A random graph from SBM(\( n, p, Q \)) is defined as follows:

- The vertex set of the graph is \( V = [n] \).
- Every vertex \( v \in V \) is independently assigned a (hidden) label \( \sigma_v \) from \([k]\) from the probability distribution \( p \) on \([k]\). That is, \( \mathbb{P}(\sigma_v = i) = p_i \) for every \( i \in [k] \).
- Given the labels of the vertices, each (unordered) pair of vertices \((u,v) \in V \times V \) is connected independently with probability \( Q_{\sigma_u, \sigma_v} \).

Example 1.1 (Symmetric communities). A simple example to keep in mind is that of symmetric communities, with more edges within communities than between communities. This is modeled by the SBM with \( p_i = 1/k \) for all \( i \in [k] \) and \( Q_{i,j} = a \) if \( i = j \) and \( Q_{i,j} = b \) otherwise, with \( a > b > 0 \).

We write \( G \sim \text{SBM}(n, p, Q) \) for a graph generated according to the SBM without the hidden vertex labels revealed. The goal of a statistical inference algorithm is to recover as many labels as possible using only the underlying graph as an observation. There are various notions of success that are worth studying.

\(^1\)Disclaimer: the literature on community detection is vast and rapidly growing. It is not our intent here to survey this literature; we refer the interested reader to the papers we cite for further references.
Figure 1: A schematic of the general stochastic block model.

- **Weak recovery** (also known as detection). An algorithm is said to *weakly recover* or *detect* the communities if it outputs a partition of the nodes which is positively correlated with the true partition, with high probability (whp).  

- **Partial recovery.** How much can be recovered about the communities? An algorithm is said to *recover communities with an accuracy of* $\alpha \in [0,1]$ if it outputs a labelling of the nodes which agrees with the true labelling on a fraction $\alpha$ of the nodes whp. An important special case is when only $o(n)$ vertices are allowed to be misclassified whp, known as *weak consistency* or *almost exact recovery*.

- **Exact recovery** (also known as *recovery* or *strong consistency*). The strongest notion of reconstruction is to recover the labels of all vertices exactly whp. When this is not possible, it can still be of interest to understand which communities can be exactly recovered, if not all; this is sometimes known as *"partial-exact-recovery".*

In all the notions above, the agreement of a partition with the true partition is maximized over all relabellings of the communities, since we are not interested in the specific original labelling per se, but rather the partition (community structure) it induces.

The different notions of recovery naturally lead to studying different regimes of the parameters. For weak recovery to be possible, many vertices in all but one community should be non-isolated (in the symmetric case this means that there should be a giant component), requiring the edge probabilities to be $\Omega(1/n)$. For exact recovery, all vertices in all but one community should be non-isolated (in the symmetric case this means that the graph should be connected), requiring the edge probabilities to be $\Omega(\ln(n)/n)$. In these regimes it is natural to scale the edge probability matrices accordingly, i.e., to consider $\text{SBM}(n,p,Q/n)$ or $\text{SBM}(n,p,\ln(n)Q/n)$, where $Q \in \mathbb{R}^{k \times k}$.

There has been lots of work in the past few years understanding the fundamental limits to recovery under the various notions discussed above. For weak recovery there is a sharp phase transition, the threshold of which was first conjectured in [8]. This was proven first for two symmetric communities [16,17] and then for multiple communities [3]. Partial recovery is less well understood, and finding the fraction of nodes that can be correctly recovered for a given set of parameters is an open problem; see [18] for results in this direction for two symmetric communities.

In this lecture we are interested in exact recovery, for which Abbe and Sandon gave the value of the threshold for the general SBM, and showed that a quasi-linear time algorithm works all the way  

---

2In these notes “with high probability” stands for with probability tending to 1 as the number of nodes in the graph, $n$, tends to infinity.
to the threshold \[2\] (building on previous work that determined the threshold for two symmetric communities [1, 19]). The remainder of this lecture is an exposition of their main results and a few of the key ideas that go into proving and understanding it.

1.2 From exact recovery to testing multivariate Poisson distributions

Recall that we are interested in the logarithmic degree regime for exact recovery, i.e., we consider \( G \sim \text{SBM}(n, p, \ln(n)Q/n) \), where \( Q \in \mathbb{R}^{k \times k} \) is independent of \( n \). We also assume that the communities have linear size, i.e., that \( p \) is independent of \( n \), and \( p_i \in (0, 1) \) for all \( i \). Our goal is to recover the labels of all the vertices whp.

As a thought experiment, imagine that not only is the graph \( G \) given, but also all vertex labels are revealed, except for that of a given vertex \( v \in V \). Is it possible to determine the label of \( v \)?

![Figure 2: Suppose all community labels are known except that of vertex \( v \). Can the label of \( v \) be determined based on its neighbors’ labels?](image)

Understanding this question is key for understanding exact recovery, since if the error probability of this is too high, then exact recovery will not be possible. On the other hand, it turns out that in this regime it is possible to recover all but \( o(n) \) labels using an initial partial recovery algorithm. The setup of the thought experiment then becomes relevant, and if we can determine the label of \( v \) given the labels of all the other nodes with low error probability, then we can correct all errors made in the initial partial recovery algorithm, leading to exact recovery. We will come back to the connection between the thought experiment and exact recovery; for now we focus on understanding this thought experiment.

Given the labels of all vertices except \( v \), the information we have about \( v \) is the number of nodes in each community it is connected to. In other words, we know the degree profile \( d(v) \) of \( v \), where, for a given labelling of the graph’s vertices, the \( i \)-th component \( d_i(v) \) is the number of edges between \( v \) and the vertices in community \( i \).

The distribution of the degree profile \( d(v) \) depends on the community that \( v \) belongs to. Recall that the community sizes are given by a multinomial distribution with parameters \( n \) and \( p \), and hence the relative size of community \( i \in [k] \) concentrates on \( p_i \). Thus if \( \sigma_v = j \), the degree profile \( d(v) = (d_1(v), \ldots, d_k(v)) \) can be approximated by independent binomials, with \( d_i(v) \) approximately distributed as \( \text{Bin} (np_i, \ln(n)Q_{i,j}/n) \), where \( \text{Bin}(m, q) \) denotes the binomial distribution with \( m \) trials and success probability \( q \). In this regime, the binomial distribution is well-approximated by a Poisson distribution of the same mean. In particular, Le Cam’s inequality gives that

\[
\text{TV} \left( \text{Bin} (na, \ln(n)b/n), \text{Poi} (ab \ln(n)) \right) \leq \frac{2ab^2 (\ln(n))^2}{n},
\]

4
where \( \text{Poi}(\lambda) \) denotes the Poisson distribution with mean \( \lambda \), and TV denotes the total variation distance. Using the additivity of the Poisson distribution and the triangle inequality, we get that

\[
\text{TV} \left( \mathcal{L}(d(v)), \text{Poi} \left( \ln(n) \sum_{i \in [k]} p_i Q_{i,j} e_i \right) \right) = O \left( \frac{(\ln(n))^2}{n} \right),
\]

where \( \mathcal{L}(d(v)) \) denotes the law of \( d(v) \) and \( e_i \) is the \( i \)-th unit vector.

Thus the degree profile of a vertex in community \( j \) is approximately Poisson distributed with mean \( \ln(n) \sum_{i \in [k]} p_i Q_{i,j} e_i \). Defining \( P = \text{diag}(p) \), this can be abbreviated as \( \ln(n) \mathcal{P}_j \), where \( (\mathcal{P}Q)_j \) denotes the \( j \)-th column of the matrix \( PQ \). We call the quantity \( (\mathcal{P}Q)_j \) the community profile of community \( j \); this is the quantity that determines the distribution of the degree profile of vertices from a given community.

Our thought experiment has thus been reduced to a Bayesian hypothesis testing problem between \( k \) multivariate Poisson distributions. The prior on the label of \( v \) is given by \( p \), and we get to observe the degree profile \( d(v) \), which comes from one of \( k \) multivariate Poisson distributions, which have mean \( \ln(n) \times \) the community profiles \( (\mathcal{P}Q)_j \), \( j \in [k] \).

### 1.3 Testing multivariate Poisson distributions

We now turn to understanding the testing problem described above; the setup is as follows. We consider a Bayesian hypothesis testing problem with \( k \) hypotheses. The random variable \( H \) takes values in \( [k] \) with prior given by \( p \), i.e., \( \mathbb{P}(H = j) = p_j \). We do not observe \( H \), but instead we observe a draw from a multivariate Poisson distribution whose mean depends on the realization of \( H \): given \( H = j \), the mean is \( \lambda(j) \in \mathbb{R}_+^k \). In short:

\[
D \mid H = j \sim \text{Poi} \left( \lambda(j) \right), \quad j \in [k].
\]

In more detail:

\[
\mathbb{P}(D = d \mid H = j) = \mathcal{P}_{\lambda(j)}(d), \quad d \in \mathbb{Z}_+^k,
\]

where

\[
\mathcal{P}_{\lambda(j)}(d) = \prod_{i \in [k]} \mathcal{P}_{\lambda_i(j)}(d_i)
\]

and

\[
\mathcal{P}_{\lambda_i(j)}(d_i) = \frac{\lambda_i(j)^{d_i}}{d_i!} e^{-\lambda_i(j)}.
\]

Our goal is to infer the value of \( H \) from a realization of \( D \). The error probability is minimized by the maximum a posteriori (MAP) rule, which, upon observing \( D = d \), selects

\[
\arg \max_{j \in [k]} \mathbb{P}(D = d \mid H = j) p_j
\]

as an estimate for the value of \( H \), with ties broken arbitrarily. Let \( P_e \) denote the error of the MAP estimator. One can think of the MAP estimator as a tournament of \( k - 1 \) pairwise comparisons of the hypotheses: if \( \mathbb{P}(D = d \mid H = i) p_i > \mathbb{P}(D = d \mid H = j) p_j \) then the MAP estimate is not \( j \). The probability that one makes an error during such a comparison is exactly

\[
P_e(i, j) := \sum_{x \in \mathbb{Z}_+^k} \min \left\{ \mathbb{P}(D = x \mid H = i) p_i, \mathbb{P}(D = x \mid H = j) p_j \right\}.
\]
For finite $k$, the error of the MAP estimator is on the same order as the largest pairwise comparison error, i.e., $\max_{i,j} P_e(i,j)$. In particular, we have that

$$\frac{1}{k-1} \sum_{i<j} P_e(i,j) \leq P_e \leq \sum_{i<j} P_e(i,j).$$  \hfill (1.2)

**Exercise 1.1.** Show (1.2).

Thus we desire to understand the magnitude of the error probability $P_e(i,j)$ in (1.1) in the particular case when the conditional distribution of $D$ given $H$ is a multivariate Poisson distribution with mean vector on the order of $\ln (n)$. The following result determines this error up to first order in the exponent.

**Lemma 1.2** (Abbe and Sandon [2]). For any $c_1, c_2 \in (0, \infty)^k$ with $c_1 \neq c_2$ and $p_1, p_2 > 0$, we have

$$\sum_{x \in \mathbb{Z}_k^+} \min \left\{ P_{\ln(n)c_1}(x) p_1, P_{\ln(n)c_2}(x) p_2 \right\} = O \left( n^{-D_+(c_1, c_2) - \frac{\ln \ln(n)}{2 \ln(n)}} \right),$$  \hfill (1.3)

$$\sum_{x \in \mathbb{Z}_k^+} \min \left\{ P_{\ln(n)c_1}(x) p_1, P_{\ln(n)c_2}(x) p_2 \right\} = \Omega \left( n^{-D_+(c_1, c_2) - \frac{k \ln \ln(n)}{2 \ln(n)}} \right),$$  \hfill (1.4)

where

$$D_+(c_1, c_2) = \max_{t \in [0,1]} \sum_{i \in [k]} \left( tc_1(i) + (1-t) c_2(i) - c_1(i)^t c_2(i)^{1-t} \right).$$  \hfill (1.5)

We do not go over the proof of this statement, which we leave to the reader as a challenging exercise.

**Exercise 1.2.** Prove Lemma 1.2.

Our conclusion is thus that the error exponent in testing multivariate Poisson distributions is given by the explicit quantity $D_+$ in (1.5). The discussion in Section 1.2 then implies that $D_+$ plays an important role in the threshold for exact recovery. In particular, it intuitively follows from Lemma 1.2 that a necessary condition for exact recovery should be that

$$\min_{i,j \in [k], i \neq j} D_+ \left( (PQ)_i, (PQ)_j \right) \geq 1.$$

Suppose on the contrary that $D_+ \left( (PQ)_i, (PQ)_j \right) < 1$ for some $i$ and $j$. This implies that the error probability in the testing problem is $\Omega \left( n^{\varepsilon-1} \right)$ for some $\varepsilon > 0$ for all vertices in communities $i$ and $j$. Since the number of vertices in these communities is linear in $n$, and most of the hypothesis testing problems are approximately independent, one expects there to be no error in the testing problems with probability at most $(1 - \Omega \left( n^{\varepsilon-1} \right))^\Omega(n) = \exp (-O (n^{\varepsilon})) = o(1)$.

### 1.4 Chernoff-Hellinger divergence

Before moving on to the threshold for exact recovery in the general SBM, we discuss connections of $D_+$ to other, well-known measures of divergence. Writing

$$D_t(\mu, \nu) := \sum_{x \in [k]} \left( t \mu(x) + (1-t) \nu(x) - \mu(x)^t \nu(x)^{1-t} \right)$$
we have that
\[ D_+ (\mu, \nu) = \max_{t \in [0,1]} D_t (\mu, \nu). \]

For any fixed \( t \), \( D_t \) can be written as
\[ D_t (\mu, \nu) = \sum_{x \in [k]} \nu(x) f_t \left( \frac{\mu(x)}{\nu(x)} \right), \]
where \( f_t(x) = 1 - t + tx - x^t \), which is a convex function. Thus \( D_t \) is an \( f \)-divergence, part of a family of divergences that generalize the Kullback-Leibler (KL) divergence (also known as relative entropy), which is obtained for \( f(x) = x \ln(x) \). The family of \( f \)-divergences with convex \( f \) share many useful properties, and hence have been widely studied in information theory and statistics. The special case of \( D_{1/2} (\mu, \nu) = \frac{1}{2} \left\| \sqrt{\mu} - \sqrt{\nu} \right\|_2^2 \) is known as the Hellinger divergence.

The Chernoff divergence is defined as \( C^* (\mu, \nu) = \max_{t \in (0,1)} -\log \sum_{x} \mu(x)^t \nu(x)^{1-t} \), and so if \( \mu \) and \( \nu \) are probability vectors, then \( D_+ (\mu, \nu) = 1 - e^{-C^*(\mu, \nu)} \). Because of these connections, Abbe and Sandon termed \( D_+ \) the Chernoff-Hellinger divergence.

While the quantity \( D_+ \) still might seem mysterious, even in light of these connections, a useful point of view is that Lemma 1.2 gives \( D_+ \) an operational meaning.

**1.5 Characterizing exact recoverability using CH-divergence**

Going back to the exact recovery problem in the general SBM, let us jump right in and state the recoverability threshold of Abbe and Sandon: exact recovery in \( \text{SBM}(n,p,\ln(n)Q/n) \) is possible if and only if the CH-divergence between all pairs of community profiles is at least 1.

**Theorem 1.3** (Abbe and Sandon [2]). Let \( k \in \mathbb{Z}_+ \) denote the number of communities, let \( p \in (0,1)^k \) with \( \|p\|_1 = 1 \) denote the community prior, let \( P = \text{diag}(p) \), and let \( Q \in (0,\infty)^{k \times k} \) be a symmetric \( k \times k \) matrix with no two rows equal. Exact recovery is solvable in \( \text{SBM}(n,p,\ln(n)Q/n) \) if and only if
\[
\min_{i,j \in [k], i \neq j} D_+ \left( (PQ)_i, (PQ)_j \right) \geq 1. \tag{1.6}
\]

This theorem thus provides an operational meaning to the CH-divergence for the community recovery problem.

**Example 1.4** (Symmetric communities). Consider again \( k \) symmetric communities, that is, \( p_i = 1/k \) for all \( i \in [k] \), \( Q_{i,j} = a \) if \( i = j \), and \( Q_{i,j} = b \) otherwise, with \( a,b > 0 \). Then exact recovery is solvable if and only if
\[
\left| \sqrt{a} - \sqrt{b} \right| \geq \sqrt{k}. \tag{1.7}
\]

We note that in this case \( D_+ \) is the same as the Hellinger divergence.

**Exercise 1.3.** Deduce from Theorem 1.3 that (1.7) gives the threshold in the example above.

**1.5.1 Achievability**

Let us now see how Theorem 1.3 follows from the hypothesis testing results, starting with the achievability. When the condition (1.6) holds, then Lemma 1.2 tells us that in the hypothesis testing problem between Poisson distributions the error of the MAP estimate is \( o(1/n) \). Thus if the setting of the thought experiment described in Section 1.2 applies to every vertex, then by looking at the degree profiles of the vertices we can correctly reclassify all vertices, and the probability...
that we make an error is $o(1)$ by a union bound. However, the setting of the thought experiment does not quite apply. Nonetheless, in this logarithmic degree regime it is possible to partially reconstruct the labels of the vertices, with only $o(n)$ vertices being misclassified. The details of this partial reconstruction procedure would require a separate lecture—in brief, it determines whether two vertices are in the same community or not by looking at how their $\log(n)$ size neighborhoods interact—so now we will take this for granted.

It is possible to show that there exists a constant $\delta$ such that if one estimates the label of a vertex $v$ based on classifications of its neighbors that are wrong with probability $x$, then the probability of misclassifying $v$ is at most $n^{\delta x}$ times the probability of error if all the neighbors of $v$ were classified correctly. The issue is that the standard partial recovery algorithm has a constant error rate for the classifications, thus the error rate of the degree profiling step could be $n^c$ times as large as the error in the hypothesis testing problem, for some $c > 0$. This is an issue when

$$\min_{i \neq j} D_+ \left( (PQ)_i, (PQ)_j \right) < 1 + c.$$ 

To get around this, one can do multiple rounds of more accurate classifications. First, one obtains a partial reconstruction of the labels with an error rate that is a sufficiently low constant. After applying the degree-profiling step to each vertex, the classification error at each vertex is now $O(n^{-c'})$ for some $c' > 0$. Hence after applying another degree-profiling step to each vertex, the classification error at each vertex will now be at most $n^{\delta \times O(n^{-c'}) \times o(1/n)} = o(1/n)$. Thus applying a union bound at this stage we can conclude that all vertices are correctly labelled whp.

### 1.5.2 Impossibility

The necessity of condition (1.6) was already described at a high level at the end of Section 1.3. Here we give some details on how to deal with the dependencies that arise.

Assume that (1.6) does not hold, and let $i$ and $j$ be two communities that violate the condition, i.e., for which $D_+ \left( (PQ)_i, (PQ)_j \right) < 1$. We want to argue that vertices in communities $i$ and $j$ cannot all be distinguished, and so any classification algorithm has to make at least one error whp. An important fact that we use is that the lower bound (1.4) arises from a particular choice of degree profile that is both likely for the two communities. Namely, define the degree profile $x$ by

$$x_\ell = \left\lfloor (PQ)^{\ell i}_{(PQ)^{1-\ell} j} \ln(n) \right\rfloor$$

for every $\ell \in [k]$, where $t \in [0,1]$ is the maximizer in $D_+ \left( (PQ)_i, (PQ)_j \right)$, i.e., the value for which $D_+ \left( (PQ)^{\ell}_i, (PQ)^{1-\ell}_j \right) = D_\ell \left( (PQ)^{\ell}_i, (PQ)^{1-\ell}_j \right)$. Then Lemma 1.2 tells us that for any vertex in community $i$ or $j$, the probability that it has degree profile $x$ is at least

$$\Omega \left( n^{-D_+(PQ)_i, (PQ)_j} / (\ln(n))^{k/2} \right),$$

which is at least $\Omega \left( n^{\varepsilon-1} \right)$ for some $\varepsilon > 0$ by assumption.

To show that this holds for many vertices in communities $i$ and $j$ at once, we first select a random set $S$ of $n/(\ln(n))^3$ vertices. Whp the intersection of $S$ with any community $\ell$ is within $\sqrt{n}$ of the expected value $p\ell n / (\ln(n))^3$, and furthermore a randomly selected vertex in $S$ is not connected to any other vertex in $S$. Thus the distribution of a vertex’s degree profile excluding connections to vertices in $S$ is essentially a multivariate Poisson distribution as before. We call a vertex in $S$ ambiguous if for each $\ell \in [k]$ it has exactly $x_\ell$ neighbors in community $\ell$ that are not in $S$. By Lemma 1.2 we have that a vertex in $S$ that is in community $i$ or $j$ is ambiguous with probability $\Omega \left( n^{\varepsilon-1} \right)$. By definition, for a fixed community assignment and choice of $S$, there
is no dependence on whether two vertices are ambiguous. Furthermore, due to the choice of the size of $S$, whp there are at least $\ln(n)$ ambiguous vertices in community $i$ and at least $\ln(n)$ ambiguous vertices in community $j$ that are not adjacent to any other vertices in $S$. These $2\ln(n)$ are indistinguishable, so no algorithm classifies all of them correctly with probability greater than $1/(2\ln(n))$, which tends to 0 as $n \to \infty$.

1.5.3 The finest exact partition recoverable

We conclude by mentioning that this threshold generalizes to finer questions. If exact recovery is not possible, what is the finest partition that can be recovered? We say that exact recovery is solvable for a community partition $[k] = \bigcup_{s=1}^{t} A_s$, where $A_s$ is a subset of $[k]$, if there exists an algorithm that whp assigns to every vertex an element of $\{A_1, \ldots, A_t\}$ that contains its true community. Abbe and Sandon show that the finest partition that is exactly recoverable can also be expressed using CH-divergence in a similar fashion. It is the largest collection of disjoint subsets such that the CH-divergence between these subsets is at least 1, where the CH-divergence between two subsets is defined as the minimum of the CH-divergences between any two community profiles in these subsets.

**Theorem 1.5 (Abbe and Sandon [2]).** Under the same settings as in Theorem 1.3, exact recovery is solvable in SBM ($n, p, \ln(n)Q/n$) for a partition $[k] = \bigcup_{s=1}^{t} A_s$ if and only if

$$D_+ \left((PQ)_i, (PQ)_j \right) \geq 1$$

for every $i$ and $j$ in different subsets of the partition.
2 Lecture 2: Estimating the dimension of a random geometric graph on a high-dimensional sphere

Many real-world networks have strong structural features and our goal is often to recover these hidden structures. In the previous lecture we studied the fundamental limits of inferring communities in the stochastic block model, a natural generative model for graphs with community structure. Another possibility is geometric structure. Many networks coming from physical considerations naturally have an underlying geometry, such as the network of major roads in a country. In other networks this stems from a latent feature space of the nodes. For instance, in social networks a person might be represented by a feature vector of their interests, and two people are connected if their interests are close enough; this latent metric space is referred to as the social space [12].

In such networks the natural questions probe the underlying geometry. Can one detect the presence of geometry? If so, can one estimate various aspects of the geometry, e.g., an appropriately defined dimension? In this lecture we study these questions in a particularly natural and simple generative model of a random geometric graph: \( n \) points are picked uniformly at random on the \( d \)-dimensional sphere, and two points are connected by an edge if and only if they are sufficiently close.

We are particularly interested in the high-dimensional regime, motivated by recent advances in all areas of applied mathematics, and in particular statistics and learning theory, where high-dimensional feature spaces are becoming the new norm. While the low-dimensional regime has been studied for a long time in probability theory [20], the high-dimensional regime brings about a host of new and interesting questions.

2.1 A simple random geometric graph model and basic questions

Let us now define more precisely the random geometric graph model we consider and the questions we study. In general, a geometric graph is such that each vertex is labeled with a point in some metric space, and an edge is present between two vertices if the distance between the corresponding labels is smaller than some prespecified threshold. We focus on the case where the underlying metric space is the Euclidean sphere \( S^{d-1} = \{ x \in \mathbb{R}^d : \|x\|_2 = 1 \} \), and the latent labels are i.i.d. uniform random vectors in \( S^{d-1} \). We denote this model by \( G(n, p, d) \), where \( n \) is the number of vertices and \( p \) is the probability of an edge between two vertices (\( p \) determines the threshold distance for connection). This model is closely related to latent space approaches to social network analysis [12].

Slightly more formally, \( G(n, p, d) \) is defined as follows. Let \( X_1, \ldots, X_n \) be independent random vectors, uniformly distributed on \( S^{d-1} \). In \( G(n, p, d) \), distinct vertices \( i \in [n] \) and \( j \in [n] \) are connected by an edge if and only if \( \langle X_i, X_j \rangle \geq t_{p, d} \), where the threshold value \( t_{p, d} \in [-1, 1] \) is such that \( \mathbb{P} (\langle X_1, X_2 \rangle \geq t_{p, d}) = p \). For example, when \( p = 1/2 \) we have \( t_{p, d} = 0 \).

The most natural random graph model without any structure is the standard Erdős-Rényi random graph \( G(n, p) \), where any two of the \( n \) vertices are independently connected with probability \( p \).

We can thus formalize the question of detecting underlying geometry as a simple hypothesis testing question. The null hypothesis is that the graph is drawn from the Erdős-Rényi model, while the alternative is that it is drawn from \( G(n, p, d) \). In brief:

\[
H_0 : G \sim G(n, p), \quad H_1 : G \sim G(n, p, d). \tag{2.1}
\]

To understand this question, the basic quantity we need to study is the total variation distance between the two distributions on graphs, \( G(n, p) \) and \( G(n, p, d) \), denoted by \( \text{TV} (G(n, p), G(n, p, d)) \);

\[4\]This lecture is based on [5].
recall that the total variation distance between two probability measures \( P \) and \( Q \) is defined as
\[
\text{TV} (P, Q) = \frac{1}{2} \| P - Q \|_1 = \sup_A |P(A) - Q(A)|.
\]
We are interested in particular in the case when the dimension \( d \) is large, growing with \( n \). It is intuitively clear that if the geometry is too high-dimensional, then it is impossible to detect it, while a low-dimensional geometry will have a strong effect on the generated graph and will be detectable. How fast can the dimension grow with \( n \) while still being able to detect it? Most of this lecture will focus on this question.

If we can detect geometry, then it is natural to ask for more information. Perhaps the ultimate goal would be to find an embedding of the vertices into an appropriate dimensional sphere that is a true representation, in the sense that the geometric graph formed from the embedded points is indeed the original graph. More modestly, can the dimension be estimated? We touch on this question at the end of the lecture.

2.2 The dimension threshold for detecting underlying geometry

The high-dimensional setting of the random geometric graph \( G(n, p, d) \) was first studied by Devroye, György, Lugosi, and Udina \[9\], who showed that if \( n \) is fixed and \( d \to \infty \), then
\[
\text{TV} (G(n, p), G(n, p, d)) \to 0,
\]
that is, geometry is indeed lost in high dimensions. More precisely, they show that this convergence happens when \( d \gg n^{7/2} \).\footnote{Throughout these notes we use standard asymptotic notation; for instance, \( f(t) \ll g(t) \) as \( t \to \infty \) if \( \lim_{t \to \infty} f(t)/g(t) = 0 \).} This follows by observing that for fixed \( n \), the multivariate central limit theorem implies that as \( d \to \infty \), the inner products of the latent vectors converge in distribution to a standard Gaussian:
\[
\left( \frac{1}{\sqrt{d}} \langle X_i, X_j \rangle \right)_{\{i,j\} \in \binom{[n]}{2}} \overset{d \to \infty}{\Rightarrow} N\left(0, I_{\binom{n}{2}}\right).
\]
The Berry-Esseen theorem gives a convergence rate, which then allows to show that for any graph \( G \) on \( n \) vertices, \( |\mathbb{P} (G(n, p) = G) - \mathbb{P} (G(n, p, d) = G)| = O\left(\sqrt{n^3/d}\right) \); the factor of \( 2^{n^2/2} \) comes from applying this bound to every term in the \( L_1 \) distance.

However, the result above is not tight, and we seek to understand the fundamental limits to detecting underlying geometry. The dimension threshold for dense graphs was recently found in \[5\], and it turns out that it is \( d \approx n^3 \), in the following sense.

**Theorem 2.1** (Bubeck, Ding, Eldan, Rácz \[5\]). Let \( p \in (0, 1) \) be fixed. Then
\[
\text{TV} (G(n, p), G(n, p, d)) \to \begin{cases} 0, & \text{if } d \gg n^3, \\ 1, & \text{if } d \ll n^3. \end{cases} \tag{2.2}
\]
Moreover, in the latter case there exists a computationally efficient test to detect underlying geometry (with running time \( O\left(n^3\right) \)).

Most of the lecture will be devoted to understanding this theorem. At the end we will consider this same question for sparse graphs (where \( p = c/n \)), where determining the dimension threshold is an intriguing open problem.
2.3 The triangle test

A natural test to uncover geometric structure is to count the number of triangles in $G$. Indeed, in a purely random scenario, vertex $u$ being connected to both $v$ and $w$ says nothing about whether $v$ and $w$ are connected. On the other hand, in a geometric setting this implies that $v$ and $w$ are close to each other due to the triangle inequality, thus increasing the probability of a connection between them. This, in turn, implies that the expected number of triangles is larger in the geometric setting, given the same edge density. Let us now compute what this statistic gives us.

Figure 3: Given that $u$ is connected to both $v$ and $w$, $v$ and $w$ are more likely to be connected under $G(n, p, d)$ than under $G(n, p)$.

For a graph $G$, let $A$ denote its adjacency matrix, i.e., $A_{i,j} = 1$ if vertices $i$ and $j$ are connected, and 0 otherwise. Then $T_G(i, j, k) := A_{i,j}A_{i,k}A_{j,k}$ is the indicator variable that three vertices $i$, $j$, and $k$ form a triangle, and so the number of triangles in $G$ is

$$T(G) := \sum_{\{i,j,k\} \in \binom{[n]}{3}} T_G(i, j, k).$$

By linearity of expectation, for both models the expected number of triangles is $\binom{n}{3}$ times the probability of a triangle between three specific vertices. For the Erdős-Rényi random graph the edges are independent, so the probability of a triangle is $p^3$, and thus we have

$$\mathbb{E} [T(G(n, p))] = \binom{n}{3} p^3.$$

For $G(n, p, d)$ it turns out that for any fixed $p \in (0, 1)$ we have

$$\mathbb{P} \left( T_{G(n, p, d)} (1, 2, 3) = 1 \right) \approx p^3 \left( 1 + \frac{C_p}{\sqrt{d}} \right)$$

for some constant $C_p > 0$, which gives that

$$\mathbb{E} [T(G(n, p, d))] \geq \binom{n}{3} p^3 \left( 1 + \frac{C_p}{\sqrt{d}} \right).$$

Showing (2.4) is somewhat involved, but in essence it follows from the concentration of measure phenomenon on the sphere, namely that most of the mass on the high-dimensional sphere is located in a band of $O \left( 1/\sqrt{d} \right)$ around the equator. We sketch here the main intuition for $p = 1/2$, which is illustrated in Figure 4.

Let $X_1$, $X_2$, and $X_3$ be independent uniformly distributed points in $S^{d-1}$. Then

$$\mathbb{P} \left( T_{G(n, 1/2, d)} (1, 2, 3) = 1 \right)$$

$$= \mathbb{P} \left( \langle X_1, X_2 \rangle \geq 0, \langle X_1, X_3 \rangle \geq 0, \langle X_2, X_3 \rangle \geq 0 \right)$$

$$= \mathbb{P} \left( \langle X_2, X_3 \rangle \geq 0 | \langle X_1, X_2 \rangle \geq 0, \langle X_1, X_3 \rangle \geq 0 \right) \mathbb{P} \left( \langle X_1, X_2 \rangle \geq 0, \langle X_1, X_3 \rangle \geq 0 \right)$$

$$= \frac{1}{4} \times \mathbb{P} \left( \langle X_2, X_3 \rangle \geq 0 | \langle X_1, X_2 \rangle \geq 0, \langle X_1, X_3 \rangle \geq 0 \right),$$

12
where the last equality follows by independence. So what remains is to show that this latter conditional probability is approximately $1/2 + c/\sqrt{d}$. To compute this conditional probability what we really need to know is the typical angle is between $X_1$ and $X_2$. By rotational invariance we may assume that $X_1 = (1, 0, 0, \ldots, 0)$, and hence $\langle X_1, X_2 \rangle = X_2(1)$, the first coordinate of $X_2$. One way to generate $X_2$ is to sample a $d$-dimensional standard Gaussian and then normalize it by its length. Since the norm of a $d$-dimensional standard Gaussian is very well concentrated around $\sqrt{d}$, it follows that $X_2(1)$ is on the order of $1/\sqrt{d}$. Conditioned on $X_2(1) \geq 0$, this typical angle gives the boost in the conditional probability that we see. See Figure 4 for an illustration.

![Figure 4: If $X_1$ and $X_2$ are two independent uniform points on the $d$-dimensional sphere $S^{d-1}$, then their inner product $\langle X_1, X_2 \rangle$ is on the order of $1/\sqrt{d}$ due to the concentration of measure phenomenon on the sphere. This then implies that the probability of a triangle in $G(n, 1/2, d)$ is $(1/2)^3 + c/\sqrt{d}$ for some constant $c > 0$.](image)

Thus we see that the boost in the number of triangles in the geometric setting is $\Theta \left( n^3/\sqrt{d} \right)$ in expectation:

$$
\mathbb{E} \left[ T \left( G(n, p, d) \right) \right] - \mathbb{E} \left[ T \left( G(n, p) \right) \right] \geq \binom{n}{3} \frac{Cp}{\sqrt{d}}
$$

To be able to tell apart the two graph distributions based on the number of triangles, the boost in expectation needs to be much greater than the standard deviation.

**Exercise 2.1.** Show that

$$\text{Var} \left( T \left( G(n, p) \right) \right) = \binom{n}{3} (p^5 - p^6) + \binom{n}{4} \binom{4}{2} (p^5 - p^6)$$

and that $\text{Var} \left( T \left( G(n, p, d) \right) \right) \leq n^4$. 

13
Exercise 2.2. Show that if
\[
|E[T(G(n,p,d))] - E[T(G(n,p))]| \gg \max \left\{ \sqrt{\text{Var}(T(G(n,p)))}, \sqrt{\text{Var}(T(G(n,p,d)))} \right\},
\]
then
\[TV(G(n,p), G(n,p,d)) \to 1.\]

Putting together Exercises 2.1 and 2.2 we see that \(TV(G(n,p), G(n,p,d)) \to 1\) if \(n^3/\sqrt{d} \gg \sqrt{n^3}\), which is equivalent to \(d \ll n^2\).

2.4 Signed triangles are more powerful

While triangles detect geometry up until \(d \ll n^2\), are there even more powerful statistics that detect geometry for larger dimensions? One can check that longer cycles also only work when \(d \ll n^2\), as do several other natural statistics. Yet it turns out that the underlying geometry can be detected even when \(d \ll n^3\).

The simple idea that leads to this improvement is to consider signed triangles. We have already noticed that triangles are more likely in the geometric setting than in the purely random setting. This also means that induced wedges (i.e., when there are exactly two edges among the three possible ones) are less likely in the geometric setting. Similarly, induced single edges are more likely, and induced independent sets on three vertices are less likely in the geometric setting. Figure 5 summarizes these observations.

![Figure 5](image_url)

Figure 5: This figure summarizes which patterns are more or less likely in the geometric setting than in the purely random setting. The signed triangles statistic reweights the different patterns with positive and negative weights.

The signed triangles statistic incorporates these observations by giving the different patterns positive or negative weights. More precisely, we define
\[
\tau(G) := \sum_{\{i,j,k\} \in \binom{[n]}{3}} (A_{i,j} - p) (A_{i,k} - p) (A_{j,k} - p).
\]

The key insight motivating this definition is that the variance of signed triangles is much smaller than the variance of triangles, due to the cancellations introduced by the centering of the adjacency matrix: the \(\Theta(n^4)\) term vanishes, leaving only the \(\Theta(n^3)\) term.

Exercise 2.3. Show that
\[
E[\tau(G(n,p))] = 0
\]
and
\[
\text{Var}(\tau(G(n,p))) = \binom{n}{3} p^3 (1 - p)^3.
\]
On the other hand it can be shown that

\[ \mathbb{E} \left[ \tau (G(n,p,d)) \right] \geq c_p n^3 / \sqrt{d}, \]

so the gap between the expectations remains. Furthermore, it can also be shown that the variance also decreases for \( G(n,p,d) \) and we have

\[ \text{Var} (\tau (G(n,p,d))) \leq n^3 + \frac{3n^4}{d}. \]

Putting everything together and using Exercise \[2.2\] for the signed triangles statistic \( \tau \), we get that \( \text{TV} (G(n,p), G(n,p,d)) \to 1 \) if \( n^3 / \sqrt{d} \gg \sqrt{n^3 + n^4/d} \), which is equivalent to \( d \ll n^3 \). This concludes the proof of \(2.3\) from Theorem \(2.1\).

### 2.5 Barrier to detecting geometry: when Wishart becomes GOE

We now turn to proving \(2.2\), which, together with \(2.3\), shows that the threshold dimension for detecting geometry is \( n^3 \). This also shows that the signed triangle statistic is near-optimal, since it can detect geometry whenever \( d \ll n^3 \).

There are essentially three main ways to bound the total variation of two distributions from above: (i) if the distributions have nice formulas associated with them, then exact computation is possible; (ii) through coupling the distributions; or (iii) by using inequalities between probability metrics to switch the problem to bounding a different notion of distance between the distributions. Here, while the distribution of \( G(n,p,d) \) does not have a nice formula associated with it, the main idea is to view this random geometric graph as a function of an \( n \times n \) Wishart matrix with \( d \) degrees of freedom—i.e., a matrix of inner products of \( n \times d \)-dimensional Gaussian vectors—denoted by \( W(n,d) \). It turns out that one can view \( G(n,p) \) as (essentially) the same function of an \( n \times n \) GOE random matrix—i.e., a symmetric matrix with i.i.d. Gaussian entries on and above the diagonal—denoted by \( M(n) \). The upside of this is that both of these random matrix ensembles have explicit densities that allow for explicit computation. We explain this connection here in the special case of \( p = 1/2 \) for simplicity; see \[5\] for the case of general \( p \).

Recall that if \( Y_1 \) is a standard normal random variable in \( \mathbb{R}^d \), then \( Y_1 / \| Y_1 \| \) is uniformly distributed on the sphere \( \mathbb{S}^{d-1} \). Consequently we can view \( G(n,1/2,d) \) as a function of an appropriate Wishart matrix, as follows. Let \( Y \) be an \( n \times d \) matrix where the entries are i.i.d. standard normal random variables, and let \( W = W(n,d) = YY^T \) be the corresponding \( n \times n \) Wishart matrix. Let \( W_{ij} = \langle Y_i, Y_j \rangle = \| Y_i \|^2 \) and so \( \langle Y_i / \| Y_i \|, Y_j / \| Y_j \| \rangle = W_{ij} / \sqrt{W_{ii} W_{jj}} \). Thus the \( n \times n \) matrix \( A \) defined as

\[ A_{i,j} = \begin{cases} 1 & \text{if } W_{ij} \geq 0 \text{ and } i \neq j \\ 0 & \text{otherwise} \end{cases} \]

has the same law as the adjacency matrix of \( G(n,1/2,d) \). Denote the map that takes \( W \) to \( A \) by \( H \), i.e., \( A = H(W) \).

In a similar way we can view \( G(n,1/2) \) as a function of an \( n \times n \) matrix drawn from the Gaussian Orthogonal Ensemble (GOE). Let \( M(n) \) be a symmetric \( n \times n \) random matrix where the diagonal entries are i.i.d. normal random variables with mean zero and variance 2, and the entries above the diagonal are i.i.d. standard normal random variables, with the entries on and above the diagonal all independent. Then \( B = H(M(n)) \) has the same law as the adjacency matrix of \( G(n,p) \). Note that \( B \) only depends on the off-diagonal elements of \( M(n) \), so in the definition of \( B \) we can replace \( M(n) \) with \( M(n,d) := \sqrt{d} M(n) + d I_n \), where \( I_n \) is the \( n \times n \) identity matrix.
We can thus conclude that

\[
\text{TV}(G(n, 1/2, d), G(n, 1/2)) = \text{TV}(H(W(n, d)), H(M(n, d))) \\
\leq \text{TV}(W(n, d), M(n, d))
\]

The densities of these two random matrix ensembles are well known. Let \( \mathcal{P} \subset \mathbb{R}^{n^2} \) denote the cone of positive semidefinite matrices. When \( d \geq n \), \( W(n, d) \) has the following density with respect to the Lebesgue measure on \( \mathcal{P} \):

\[
f_{n,d}(A) := \frac{(\det(A))^{\frac{1}{2}(d-n-1)} \exp \left(-\frac{1}{2} \text{Tr}(A)\right)}{2^{\frac{1}{2}dn} \pi^{\frac{n(n-1)}{2}} \prod_{i=1}^{n} \Gamma \left(\frac{1}{2}(d+1-i)\right)},
\]

where \( \text{Tr}(A) \) denotes the trace of the matrix \( A \). It is also known that the density of a GOE random matrix with respect to the Lebesgue measure on \( \mathbb{R}^{n^2} \) is

\[
f(A) := \frac{\exp \left(-\frac{1}{4} \text{Tr}(A^2)\right)}{(2\pi)^{\frac{n(n+1)}{2}} 2^{\frac{n^2}{2}}},
\]

and so the density of \( M(n, d) \) with respect to the Lebesgue measure on \( \mathbb{R}^{n^2} \) is

\[
g_{n,d}(A) := \frac{\exp \left(-\frac{1}{4} \text{Tr}((A - dI_n)^2)\right)}{(2\pi d)^{\frac{n(n+1)}{2}} 2^{\frac{n^2}{2}}}.
\]

These explicit formulas allow for explicit calculations. In particular, one can show that the log-ratio of the densities is \( o(1) \) with probability \( 1 - o(1) \) according to the measure induced by \( M(n, d) \). This follows from writing out the Taylor expansion of the log-ratio of the densities and using known results about the empirical spectral distribution of Wigner matrices (in particular that it converges to a semi-circle law). The outcome of the calculation is the following result, proven independently and simultaneously by Bubeck et al. and Jiang and Li.

**Theorem 2.2** (Bubeck, Ding, Eldan, Rácz [5]; Jiang, Li [14]). Define the random matrix ensembles \( W(n, d) \) and \( M(n, d) \) as above. If \( d/n^3 \to \infty \), then

\[
\text{TV}(W(n, d), M(n, d)) \to 0.
\]

We conclude that it is impossible to detect underlying geometry whenever \( d \gg n^3 \).

### 2.6 Estimating the dimension

Until now we discussed detecting geometry. However, the insights gained above allow us to also touch upon the more subtle problem of estimating the underlying dimension \( d \).

Dimension estimation can also be done by counting the “number” of signed triangles as in Section 2.4. However, here it is necessary to have a bound on the difference of the expected number of signed triangles between consecutive dimensions; the lower bound of (2.5) is not enough. Still, we believe that the right hand side of (2.5) should give the true value of the expected value for an appropriate constant \( c_p \), and hence we expect to have that

\[
\mathbb{E}[\tau(G(n, p, d))] - \mathbb{E}[\tau(G(n, p, d + 1))] = \Theta \left(\frac{n^3}{d^{3/2}}\right).
\]

Thus, using the variance bound in (2.6), we get that dimension estimation should be possible using signed triangles whenever \( n^3/d^{3/2} \gg \sqrt{n^3 + n^4}/d \), which is equivalent to \( d \ll n \).

Showing (2.7) for general \( p \) seems involved; Bubeck et al. showed that it holds for \( p = 1/2 \), which can be considered as a proof of concept. We thus have the following.
Theorem 2.3 (Bubeck, Ding, Eldan, Rácz [5]). There exists a universal constant $C > 0$ such that for all integers $n$ and $d_1 < d_2$, one has

$$\text{TV} \left( G(n, 1/2, d_1), G(n, 1/2, d_2) \right) \geq 1 - C \left( \frac{d_1}{n} \right)^2.$$ 

This result is tight, as demonstrated by a result of Eldan [10], which states that when $d \gg n$, the Wishart matrices $W(n, d)$ and $W(n, d + 1)$ are indistinguishable. By the discussion in Section 2.5, this directly implies that $G(n, 1/2, d)$ and $G(n, 1/2, d + 1)$ are indistinguishable.

Theorem 2.4 (Eldan [10]). There exists a universal constant $C > 0$ such that for all integers $n < d$,

$$\text{TV} \left( G(n, 1/2, d), G(n, 1/2, d + 1) \right) \leq \text{TV} \left( W(n, d), W(n, d + 1) \right) \leq C \sqrt{\frac{d + 1}{d - n}} - 1.$$ 

2.7 The mysterious sparse regime

The discussion so far has focused on dense graphs, i.e., assuming $p \in (0, 1)$ is constant, where Theorem 2.1 tightly characterizes when the underlying geometry can be detected. The same questions are interesting for sparse graphs as well, where the average degree is constant or slowly growing with $n$. However, since there are so few edges, this regime is much more challenging.

It is again natural to consider the number of triangles as a way to distinguish between $G(n, c/n)$ and $G(n, c/n, d)$. A calculation shows that this statistic works whenever $d \ll \log_3(n)$.

Theorem 2.5 (Bubeck, Ding, Eldan, Rácz [5]). Let $c > 0$ be fixed and assume $d/\log_3(n) \to 0$. Then

$$\text{TV} \left( G \left( n, \frac{c}{n} \right), G \left( n, \frac{c}{n}, d \right) \right) \to 1.$$ 

In contrast with the dense regime, in the sparse regime the signed triangle statistic $\tau$ does not give significantly more power than the triangle statistic $T$. This is because in the sparse regime, with high probability, the graph does not contain any 4-vertex subgraph with at least 5 edges, which is where the improvement comes from in the dense regime.

The authors also conjecture that $\log_3(n)$ is the correct order where the transition happens.

Conjecture 2.6 (Bubeck, Ding, Eldan, Rácz [5]). Let $c > 0$ be fixed and assume $d/\log_3(n) \to \infty$. Then

$$\text{TV} \left( G \left( n, \frac{c}{n} \right), G \left( n, \frac{c}{n}, d \right) \right) \to 0.$$ 

The main reason for this conjecture is that, when $d \gg \log_3(n)$, $G(n, c/n)$ and $G(n, c/n, d)$ seem to be locally equivalent; in particular, they both have the same Poisson number of triangles asymptotically. Thus the only way to distinguish between them would be to find an emergent global property which is significantly different under the two models, but this seems unlikely to exist. Proving or disproving this conjecture remains a challenging open problem. The best known bound is $n^3$ from (2.2) (which holds uniformly over $p$).
References


