Theory Group Seminar Notes

Rishit Dagli

October 2022

Contents

Introduction

These are my notes for the seminars that happen in the Theory Group at The University of Toronto. Many thanks to Professor Allan Borodin for allowing me to attend the Theory Group seminars and helping out.

A PDF of these notes is available at https://rishit-dagli.github.io/cs-theory-notes/main.pdf. An online version of these notes are available at https://rishit-dagli.github.io/cs-theory-notes.

The Theory Group focuses on theory of computation. The group is interested in using mathematical techniques to understand the nature of computation and to design and analyze algorithms for important and fundamental problems.

The members of the theory group are all interested, in one way or another, in the limitations of computation: What problems are not feasible to solve on a computer? How can the infeasibility of a problem be used to rigorously construct secure cryptographic protocols? What problems cannot be solved faster using more machines? What are the limits to how fast a particular problem can be solved or how much space is needed to solve it? How do randomness, parallelism, the operations that are allowed, and the need for fault tolerance or security affect this?

1 Lower Bounds for Locally Decodable Codes from Semirandom CSP Refutation

7th October 2022

The related paper: Combinatorial lower bounds for 3-query LDCs by Alrabiah et al. [1]. Seminar by Peter Manohar. [2] [3]

1.1 Abstract

A code C is a q-locally decodable code (q-LDC) if one can recover any chosen bit b_i of the k-bit message b with good confidence by randomly querying the n-bit encoding x on at most q coordinates. Existing constructions of 2-LDCs achieve blocklength $n = exp(O(k))$, and lower bounds show that this is in fact tight. However, when $q = 3$, far less is known: the best constructions have $n = subexp(k)$, while the best known lower bounds, that have stood for nearly two decades, only show a quadratic lower bound of $n \geq \Omega(k^2)$ on the blocklength.

In this talk, we will survey a new approach to prove lower bounds for LDCs using recent advances in refuting semirandom instances of constraint satisfaction problems. These new tools yield, in the 3-query case, a near-cubic lower bound of $n \geq \tilde{\Omega}(k^3)$, improving on prior work by a polynomial factor in k.

1.2 Locally Decodable Codes

Take codes $b \in 0,1^k \to x \in 0,1^n$
Codes x are read by the decoder, $i \in [k]$, $\hat{b_i} \in 0,1$

Definition 1. *C is a (q, δ, ϵ)-locally decodable if for any x with $\triangle(x, Enc(b)) \leq \delta n$, $Dec^x(i) = b_i$ w.p. $\geq \frac{1}{2} + \epsilon$ for any i.*

Ask the question, what is the best possible rate for a q-LDC given a q?

q	Lower Bound	Upper Bound
2	$2^{\Omega(k)} \leq n$	$n \leq 2^k$
3	$k^2 \leq n$	$n \leq exp(k^{o(1)})$
$O(1)$, even	$k^{\frac{q}{q+1}} \leq n$	$n \leq exp(k^{o(1)})$
$O(1)$, odd	$k^{\frac{q+1}{q-1}} \leq n$	$n \leq exp(k^{o(1)})$

Focus on the case $q = 3$, we have gotten better bounds:

$$k \leq n \leq 2^k \tag{1}$$

$$k^2 \leq n \leq exp(exp(\sqrt{\log k \log\log k}))$$

In [1], they show that a better minimum bound can be found than these existing ones for $q = 3$:

$$k^3 \leq n \tag{2}$$

The main result is that:

Theorem 1. *Let C be a $(3, \delta, \epsilon)$-locally decodable codes. Then $n \geq \tilde{\Omega}_{\delta,\epsilon}(k^3)$.*

Semi-random CSP refutation comes to our aid to prove this! The intuitive way to put this theorem is that q-LDC lower bound is same as refuting "LDC" q-XOR.

1.3 How to prove the Theorem

The idea:

- q-LDC lower bound is same as refuting "LDC" q-XOR
  - CSP Refutation
- Proof of existing q-LDC lower bound for q even
- Proof sketch of k^3 lower bound

1.4 Normally Decodable Codes

We can see that the decoder we have can arbitrary but WLOG we can assume there are q-unif hypergraphs $H_1, H_2, \cdots H_k$ where every H_i is such that:

$$H_i \subseteq \binom{[n]}{q}$$

We can also see that:
Each H_i is a matching such that $|H_i| \geq \delta n$
and, $Dec(i)$ picks $C \leftarrow H_i$ and outputs $\sum_{j \in C} x_j$

One such example is the Hadmard code:

$$b \in 0, 1^k \mapsto f = (\langle b, v \rangle)_{v \in 0,1}^k \tag{3}$$

$$b_i = f(e_i) = f(v) + f(v + e_i)$$

Can think of this as v and $v + e_i$ are connected.

Matching vector codes are $\approx \mathbb{Z}_m^h$

1.5 Proof: Going from LDC to XOR

We suppose that our code is linear and that there exists q-unif hypergraphs $H_1, H_2, \cdots H_k$.

We also know that:
Each H_i is a matching such that $|H_i| \geq \delta n$
and, $Dec(i)$ picks $C \leftarrow H_i$ and outputs $\sum_{j \in C} x_j$

So, we start by considering a q-XOR instance ψ_b:

$$\textit{Vars: } \{x_j\}_{j \in [n]}$$
$$\textit{Over Equations: } \sum_{j \in C} x_j = b_i, \forall i \in [k], C \in H_i$$

We can write down the maximum fraction of satisfiable constraints: $val(\psi_b) = 1$ for any $b \in 0, 1^k$.

It is sufficient now if we can argue that ψ_b is unsat with high probability for some random b when $n \ll k^{\frac{q}{q-2}}$.

Now we need to refute XOR, there are many ways to argue unsatisfiability of an XOR instance. One reason why we can not use probablistic approaches here is that ψ_b only has k bits of randomness.

One way we can have some success here is to use a refutation algorithm

$$\psi \to A \to algval(\psi)$$

With this the guarantee then would be $val(\psi) \leq algval(\psi)$ which is similar to saying that if $algval(\psi) < 1$ then A refutes ψ. The ideal goal would be to refute random ψ with m constraints with high probability

However, we take a look at semi-random XOR. Our refutation algorithm and the guarantee will still be the same:

$$\psi \to A \to algval(\psi)$$

with the guarantee that $val(\psi) \leq algval(\psi)$.

So, now we generate semi-random $\psi w/m$ constraints:

- The worst case would be random q-unif hypergraph
- Random RHS b_c for each $C \in H$

The equation we have is:

$$\sum_{j \in C} x_j = b_c \tag{4}$$

And we also already know that

$$\psi_b \ is \sum_{j \in C}$$

And, $xj = b_i, i \in [k], C \in H_i$.
ψ_b is almost semi-random.

Thus, we have shown 1.3 Part 1 of Proof.

1.6 Proof: Existing q-LDC lower bound for q even

q-LDC XOR instance ψ_b is encoded by:

- q-uniform hypergraph matchings $\{H_1 \cdots H_k\}$
- right-hand sides are random $b_i \in \{\pm 1\}$
- We have constraints $\prod_{j \in C} x_j = b_i$ for all i and $C \in H_i$

We now have a goal to argue that ψ_b unsat with high probability for random when b when $n \ll k^{q/(q-2)}$

frac. constraints satisfied by $x \in \{\pm 1\}^n$ is $\frac{1}{2} + \frac{f(x)}{2}$.

Here $f(x)$ is:

$$f(x) = \frac{1}{m} \sum_i b_i \sum_{C \in H_i} \prod_{j \in C} x_j \tag{5}$$

$$m = k \cdot \delta n$$

This makes our goal to be to certify with high probability that:

$$\max_{x \in \{\pm 1\}^n} f(x) < 1 \ when \ n \ll k^{\frac{q}{q-2}} \tag{6}$$

We will now try to refute ψ_b. With Equation 5 and Equation 6 to refute ψ_b is like showing:

$$w.h.p. \quad \max_{x \in \{\pm 1\}^n} f(x) < 1 \ where \ f(x) = \frac{1}{m} \sum_i b_i \sum_{C \in H_i} \prod_{j \in C} x_j \tag{7}$$

$when \ n \ll k^{\frac{q}{q-2}}$.

The idea is to design a matrix $A \in \mathbb{R}^{N \times N}$ so that:

$$f(x) \leq ||A||_{\infty \to 1} = \max_{z,w \in \{\pm 1\}^N} z^T A w$$

As shown by Wein et al. [4] the matrix A can be indexed by

$$S \in \binom{[n]}{l}$$

Assign $x \mapsto y$ such that $y^T A y \propto f(x)$
and $y_s := \prod_{j \in S} x_j$ which is simply the tensor product.

We need to now be able to answer how to set $A(S,T)$

$$y^T A y = \sum_{S,T} y_S y_T A(S,T) = \sum_{S,T} A(S,T) \prod_{j \in S \oplus T} x_j \tag{8}$$

Which shows that we are actually using symmetric difference here.

We say that if $S \oplus T = C \in h_i$ then $\prod_{j \in S \oplus T} x_j = b_i$
$\implies A(S,T) = b_i$ if $S \oplus T = C \in H_i$

$$y^T A y = \sum_{i=1}^{k} b_i \sum_{C \in h_i} \sum_{S \oplus t = C} \prod_{j \in C} x_j = Dmf(x) \tag{9}$$

Here $D =$ number of S, T where $S \oplus T = C$.

Simplifying an earlier statement we can also say from here that: $A_C(S,T) = 1$ if $S \oplus T = C$.

For which $A_i = \sum_{C \in h_i} A_C$ and $A = \sum_{i=1}^{k} b_i A_i$

Set $y_S := \prod_{j \in S} x_j$

$y^T A y = Dmf(x) \implies Dmf(x) \leq ||A||_{\infty \to 1}$

Note that the way we defined D here it only depends on $|C| = q$, we can say:

$$D = \binom{q}{\frac{q}{2}} \binom{n-q}{l-\frac{q}{2}}$$

Also we know $A_c \in \mathbb{R}^{N \times N}$ and $N = \binom{n}{l}$.

We have already proven that $||A||_{\infty \to 1} \geq Dm \max_x f(x) \geq Dm \geq D\delta nk$

It is also interesting to note that $||A||_{\infty\to 1} \leq N||A||_2$ and we still need to be able to show that with high probability that $||A||_{\infty\to 1}$ is not too large.

Matrix Bernstein: with high probability over b_i, $||A||_2 \leq \triangle\sqrt{kl}$ where $\triangle$ is the maximum number of 1's in a row in any A_i.

Expected number of 1's per row is $\delta n\frac{D}{N} \sim n(\frac{l}{n})^{q/2}$.

We can optimistically suppose that $\triangle \sim n(\frac{l}{n})^{q/2}$ however this also needs $l \geq n^{1-2/q}$.

Then $D \cdot \delta nk \leq ||A||_{\infty\to 1} \leq N\triangle\sqrt{kl}$

$\implies k \leq l$ since $\triangle \sim \delta n\frac{D}{N}$

Now take $l = n^{1-2/q} \implies k^{q/(q-2)} \leq n$

So, $\triangle = \frac{2l}{q}$

Because H_i are matchings, a random row will have only $\approx \frac{\delta nD}{N}$ 1's.

The idea now is to prune off all the bad rows or columns in A to get B such that:

$$||A||_{\infty\to 1} \leq ||B||_{\infty\to 1} + o(N)$$

And, $\triangle_B \sim \delta n(\frac{l}{n})^{q/2}$

And now we can just use B instead which will prove q-LDC lower bound for q even.

1.7 Proof: k^3 lower bound

Recall, q-LDC XOR instance ψ_b is encoded by:

- q-uniform hypergraph matchings $\{H_1 \cdots H_k\}$
- right-hand sides are random $b_i \in \{\pm 1\}$
- We have constraints $\prod_{j\in C} x_j = b_i$ for all i and $C \in H_i$

The goal is argue that ψ_b is unsatisfiable with high probability for random b. And the idea is to design a matrix $A \in \mathbb{R}^{N\times N}$ so that:

$$f(x) \leq ||A||_{\infty\to 1} = \max_{z,w\in\{\pm 1\}^N} z^T Aw$$

The previous approach fails because the A from before requires q to be even.

One attempt is to represent rows as $|S| = l$ and columns as $|T| = l+1$. However this will only get us to $k \leq \sqrt{n}$.

We need to derive more constraints, using $C_i \oplus C_j$ get us to nk constraints so each n_j is in $\approx k$ constraints $\implies$ new nk^2 constraints.

The matrix A is indexed by S, $A(S,T) = b_i b_j$. The calculation is now:

$$nk^2 D \leq ||A||_{\infty \to 1} \leq N \triangle \sqrt{kl}$$

An optimist approach is $\triangle \sim Nk\frac{D}{N} = nk(\frac{l}{n})^2$

$\implies l \geq \sqrt{\frac{n}{k}}$

$\implies k \leq n \implies k^3 \leq n$

The row pruning tricks would still work provided that any $\{u, v\}$ is in at most $polylog(n)$ constraints.

1.8 Conclusion

This proof for $q = 3$ is not generalizable for all odd q and neither is a reduction to 2-LDC. This is particularly true because of the row pruning step.

2 Algorithms for the ferromagnetic Potts model on expanders

14th October 2022

The related paper: Algorithms for the ferromagnetic Potts model on expanders by Carlson et al. [5]. Seminar by Aditya Potukuchi.

2.1 Abstract

The ferromagnetic Potts model is a canonical example of a Markov random field from statistical physics that is of great probabilistic and algorithmic interest. This is a distribution over all 1-colorings of the vertices of a graph where monochromatic edges are favored. The algorithmic problem of efficiently sampling approximately from this model is known to be #BIS-hard, and has seen a lot of recent interest. I will outline some recently developed algorithms for approximately sampling from the ferromagnetic Potts model on d-regular weakly expanding graphs. This is achieved by a significantly sharper analysis of standard "polymer methods" using extremal graph theory and applications of Karger's algorithm to count cuts that may be of independent interest. I will give an introduction to all the topics that are relevant to the results.

2.2 The Ferromagnetic Potts Model

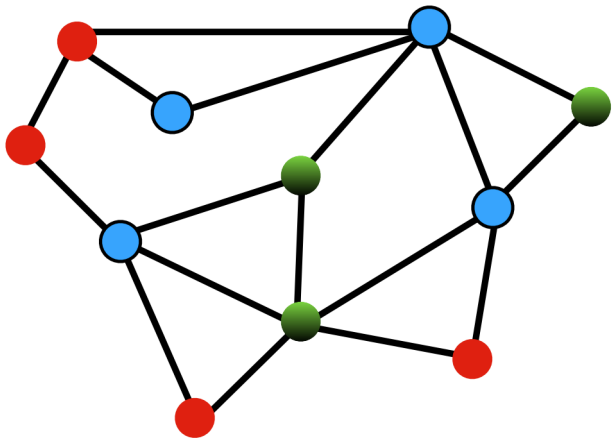

Figure 1: A sample graph

We start by defining some basic notation:

- G: finite graph on vertices V
- $q \in \mathbb{N}$, we are interested in q-colourings of the vertices in G
- $m(\chi)$: number of monochromatic edges induced by a colouring χ
- Distribution on colourings given by $p(\chi) \propto exp(\beta \cdot m(\chi))$
- $\beta \in \mathbb{R}$: parameter, inverse temperature

Notice that for $\beta < 0$ it means that we take the antiferromagnetic case. Here we talk more about when $\beta > 0$ meaning it is ferromagnetic.

This could have quite some applications:

- Modelling: Social networks, physics, chemistry, etc
- Markov Random field: Probabilistic Inference
- Connection to UGC Coulson et al. [6]
- and more.

2.3 The Problem

we know $p(\chi) \propto exp(\beta \cdot m(\chi))$

Now for $\beta = 0$ it means that we are doing a uniform q-coloring of V

For $\beta = -\infty$ we do a uniform proper coloring of G

What we need to do is given G and β, efficiently sample a coloring from this distribution.

$$p(\chi) = \frac{exp(\beta m(\chi))}{\sum_{\chi} exp(\beta m(\chi))} \tag{10}$$

We add the normalizing factor here:

$$\textit{Nomalizing factor } = \sum_{\chi} exp(\beta m(\chi))$$

Now we can also say,

$$\sum_{\chi} exp(\beta m(\chi)) =: Z_G(q, \beta) \tag{11}$$

A partition function of the model/distribution is very important for this POV.

Our problem is that given G and β we want to efficiently sample a color distribution. We give 2 facts:

1. It is enough to compute $Z_G(q, \beta)$
2. #P-hard

We now modify the problem as: Given G and β, efficiently sample **approximately** a colouring from this distribution.

ϵ approximation will have us sample a law from q such that $||p - q||_{TVD} \leq \epsilon$, thus

$$||p - q||_{TVD} := \frac{1}{2} \sum_{\chi} |p(\chi) - q(\chi)| \tag{12}$$

We modify our original problem template to now be: Given G and β, efficiently sample ϵ-**approximately** a colouring from this distribution.

Fully Polynomial Almost Uniform Sampler can allow us to sample ϵ-approximately in $poly(G, \frac{1}{\epsilon})$ time.

Instead Fully Polynomial Time Approximation Scheme: $1 \pm \epsilon$-factor approximation in $poly(G, \frac{1}{\epsilon})$ time.

We can also show for a fact that $FPTAS \iff FPAUS$.

2.4 Antiferromagnetic Potts model

The Antiferromagnetic Potts model:

$$p(\chi) \propto exp\beta \cdot m(\chi) \tag{13}$$

where $\beta < 0$

Given G and $\beta < 0$, we want to be able to give an FPAUS for this distribution. It is then equivalent to instead work on the problem: given G and $\beta < 0$, give an FPTAS for its partition function $Z_G(q, \beta)$.

From some previous work, we know that there exists a β_c such that:

- for $\beta < \beta_c$, FPTAS exists
- For $\beta < \beta_c$, no FPTAS unless $NP = RP$

We can say that this is #BIS-hard (bipartite independent sets). Thus, doing this is at least as hard as an FPTAS for the number of independent sets in bipartite graphs. If our graph has no bipartiteness then this becomes a NP-hard problem.

For now, let's consider the problem given a bipartite graph G, design an FPTAS for the number of individual sets in G. This accurately captures the difficulty of: the number of proper q-colorings of a bipartite graph for $q \geq 3$, the number of stable matchings, the number of antichains in posets.

2.5 Main Results

For our purposes we assume that G is always a d-regular graph on n vertices. Now for a set $S \subset V$, we define it's edge boundary as:

$$\nabla(S) := \#(uv \in G | u \in S, v \notin S)$$

Now, G is an η expander if for every $S \subset V$ of size at most $n/2$, we have $|\nabla(S)| \geq \eta|S|$. For example we can take a discrete cube Q_d with vertices $\{0,1\}^d$, uv is an edge if u and v differ in exactly 1 coordinate.
Using a simplification of the Harper's Theorem we can say that Q_d is a 1-expander [7].

Theorem 2. *For each $\epsilon > 0$ and there is a $d = d(\epsilon)$ and $q = q(\epsilon)$ such that there is an FPTAS for $Z_G(q, \beta)$ where G is a d-regular 2-expander providing the following conditions hold:*

- $q = poly(d)$
- $\beta \notin (2 \pm \epsilon)\frac{ln(q)}{d}$

The main result shown was that

Theorem 3. *For each $\epsilon > 0$, and d large enough, there is an FPTAS for $Z_G(q, \beta)$ where G for the class of d-regular triangle-free 1-expander grpahs providing the following conditions hold:*

- $q \geq poly(d)$
- $\beta \notin (2 \pm \epsilon)\frac{ln(q)}{d}$

This was previously known for:

- Stronger expansion and $d = q^{\Omega(d)}$
- Higher temperature and $q = d^{\Omega(d)}$

Something to note here is that $q \geq poly(d)$ should not be a necessary condition.

As well as as in the case $\beta \leq (1 - \epsilon)\beta_0$ does not require expansion or even that $q \geq poly(d)$.

2.6 Potts Distribution

We first write the order-disorder threshold of the ferromagnetic Potts model

$$\begin{aligned} \beta_0 :=& ln(\frac{q-2}{(q-1)^{1-2/d}-1}) \\ \beta_0 =& 2\frac{lnq}{d}(1 + O(\frac{1}{q})) \end{aligned} \tag{14}$$

We want to be able to know more about how the Potts distribution looks for $\beta < (1-\epsilon)\beta_0$ and for $\beta > (1+\epsilon)\beta_0$

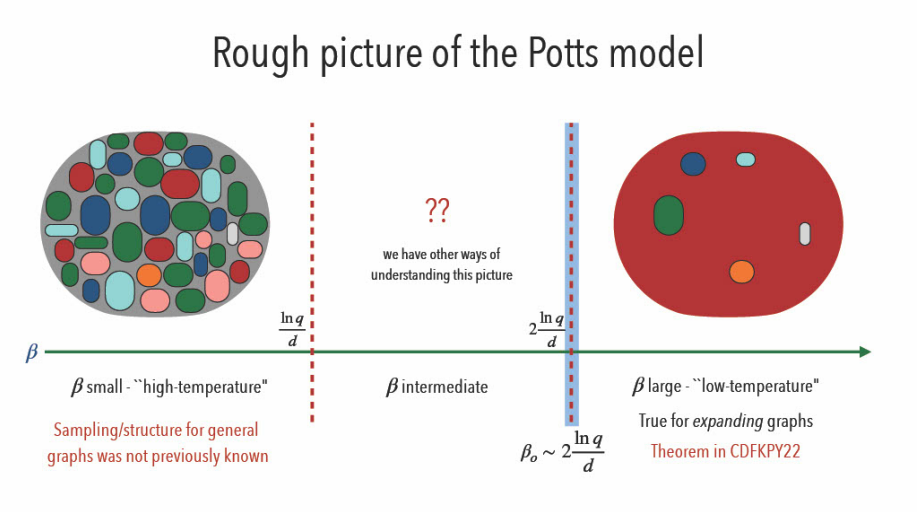

Figure 2: Rough picture of the Potts model

2.7 Results

Another result we have is:

Theorem 4. *For each $\epsilon > 0$, let d be large enough $q \geq poly(d)$, and G be a d-regular 2-expander graph on n vertices then,*

- *For $\beta < (1-\epsilon)\beta_0$, every colour class has size $n/q(1 \pm o(1))$ with high probability*
- *For $\beta > (1+\epsilon)\beta_0$, every colour class has size $n - o(n)$ with high probability*

The strategy we have, to prove the theorem for $\beta < (1-\epsilon)\beta_0$:

- Pass to the Random Cluster Model
- Distribution on subsets of edges: $p(A) \propto q^{k(A)}(e^\beta - 1)^{|A|}$
- $Z_G^{RC}(q,\beta) = Z_G^{Potts}(q,\beta)$
- Sampling algorithm: Sample from random cluster model, give each connected component a uniform color
- Standard polymer methods + careful enumeration

2.8 Polymer Methods

The motivating idea is to visualize the state for β large at low temperature as ground state + defects.

Typical Colouring = Ground State + Defects

Polymer methods are pretty useful in such cases. These were first proposed in [8] and originated in statistical physics. We take G to be our defect graph and each node in this represents a defect.

Now using Polymer methods $X \sim_G Y$

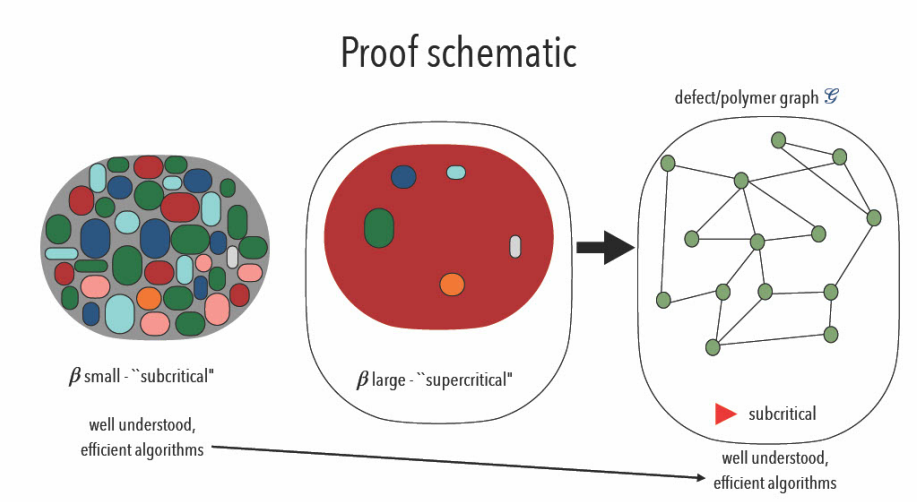

Figure 3: Proof Schematic

Ideas is to $Z_G(q,\beta) \sim Z_{red} + Z_{blue} + \dots$ where $Z_{red} \approx e^{\beta nd/2}$

$Z_{red}e^{-\beta nd/2} = \sum_{I \subset V(G)} \prod_{\gamma \in I} w_\gamma$ where w_γ is the weight of polymer γ.

We now move towards cluster expansion: multivariate in the w_γ Taylor expansion of:

$$ln(\sum_{I \subset V(G)} \prod_{\gamma \in I} w_\gamma)$$

This is an infinite sum, so convergence is not guaranteed however convergence can be established by verifying the Kotecký-Preiss criterion.

We also want to answer how many connected subsets are there of a given edge boundary in an η-expander?

A heuristic we have is to count the number of such subsets that contain a given vertex u: a typical connected subgraph of size a is tree-like, i.e., has edge boundary $a \cdot d$.

Working backward, a typically connected subgraph with edge boundary size b has $O(b/d)$ vertices. The number of such subgraphs $\leq$ number of connected subgraphs of size $O(b/d)$ containing u. The original number of subsets is also $\leq$ Number of rooted (at u) trees with $O(b/d)$ vertices and maximum degree at most $d = d^{O(b/d)}$. Thus,

Theorem 5. *At most $d^{O(1+1/\eta)b/d}$ connected subsets in an η expander that contains u have edge boundary of size at most b.*

Another question to ask is how many q-colorings of an η-expander induce at most k non-monochromatic edges?

Easiest way is to make k non-monochrimatic edges is to color all but k/d randomly chosen vertices with the same color. Now, k small $\implies$ these vertices likely form an independent set. we now color these k/d vertices arbitrarily. There are:

$$\binom{n}{k/d} q^{k/d+1}$$

ways.

Theorem 6. *For η-expanders and $q \geq poly(d)$ there are at most $n^4 q^{O(k/d)}$ possible colourings.*

Now we also know the maximum value of $Z_G(q, \beta)$ over all graphs G with n vertices, m edges, and max degree d. This will always be attained when G is a disjoint union of K_{d+1} and K_1

3 Statistical Learning using Compression

18th October 2022

The related paper: Adversarially Robust Learning with Tolerance by Ashtiani et al. [9]. Seminar by Hassan Ashtiani.

3.1 Abstract

Characterizing the sample complexity of different machine learning tasks is one of the central questions in statistical learning theory. For example, the classic Vapnik-Chervonenkis theory characterizes the sample complexity of binary classification. Despite this early progress, the sample complexity of many important learning tasks — including density estimation and learning under adversarial perturbations — are not yet resolved. In this talk, we review the less conventional approach of using compression schemes for proving sample complexity upper bounds, with specific applications in learning under adversarial perturbations and learning Gaussian mixture models.

3.2 Some background

We start by defining some notation:

- Z: domain set
- D_Z: distribution over Z
- S: i.i.d sample from D_Z
- H: class of models/hypotheses
- $L(D_Z, H) \to \mathbb{R}$: loss/ error function
- $OPT = inf_{h \in H} L(D_z, h)$: best achievable
- $A_{Z,H} : Z^* \to H$: learner

3.3 Density Estimation

Our goal is that for every D_Z, $A_{Z,H}(S)$ we want it to be comparable to OPT with high probability.

We take the example of density estimation in this case $L(D_Z, h) = d_{TV}(D_z, h)$. Now, $A_{Z,H}$ probably approximately correct learns H with $m(\epsilon, \delta)$ samples if for all D_Z and for all ϵ with a $\delta \in (0, 1)$. Now if $S \sim D_Z^{m(\epsilon,\delta)}$ then:

$$\Pr_S[L(D_Z, A_{Z,H}(S)) > \epsilon + C \cdot OPT] < \delta \tag{15}$$

Now if we take the example of $C = 2$, let H be the set of all Gaussians in $\mathbb{R}^d$ then:

$$m(\epsilon, \delta) = O\left(\frac{d^2 + \log 1/\delta}{\epsilon^2}\right)$$

We will now modify the above equation. Now, $A_{Z,H}$ probably approximately correct learns H with $m(\epsilon)$ samples if for all D_Z and for all $\epsilon \in (0, 1)$. Now if $S \sim D_Z^{m(\epsilon)}$ then:

$$\Pr_S[L(D_Z, A_{Z,H}(S)) > \epsilon + C \cdot OPT] < 0.01 \tag{16}$$

For the example of $C = 2$, let H be the set of all Gaussians in $\mathbb{R}^d$ then:

$$m(\epsilon, \delta) = O\left(\frac{d^2}{\epsilon^2}\right)$$

3.4 Binary Classification (with adv. pertubations)

For the example of binary classification, we have $Z = X \times \{0, 1\}$ and h is some model which maps from $h : X \to \{0, 1\}$.

We also have $l(h, x, y) = 1h(x) \neq y$ and then we will have the L be $L(D_Z, h) = E_{(x,y)\sim D_Z} l(h, x, y)$.

Now, $A_{Z,H}$ probably approximately correct learns H with $m(\epsilon)$ samples if for all D_Z and for all $\epsilon \in (0, 1)$. Now if $S \sim D_Z^{m(\epsilon)}$ then:

$$\Pr_S[L(D_Z, A_{Z,H}(S)) > \epsilon + C \cdot OPT] < 0.01 \tag{17}$$

Now H is the set of all half spaces in $\mathbb{R}^d$ then:

$$m(\epsilon) = O\left(\frac{d}{\epsilon^2}\right)$$

For the example of binary classification, we have $Z = X \times \{0, 1\}$ and h is some model which maps from $h : X \to \{0, 1\}$.

We also have $l^U(h, x, y) =$ adversarial pertubations and then we will have the L^U be $L^U(D_Z, h) = E_{(x,y)\sim D_Z} l^U(h, x, y)$.

Now, $A_{Z,H}$ probably approximately correct learns H with $m(\epsilon)$ samples if for all D_Z and for all $\epsilon \in (0, 1)$. Now if $S \sim D_Z^{m(\epsilon)}$ then:

$$\Pr_S[L^U(D_Z, A_{Z,H}(S))\epsilon + OPT] < 0.01 \tag{18}$$

Now, consider the following H_1 and H_2:

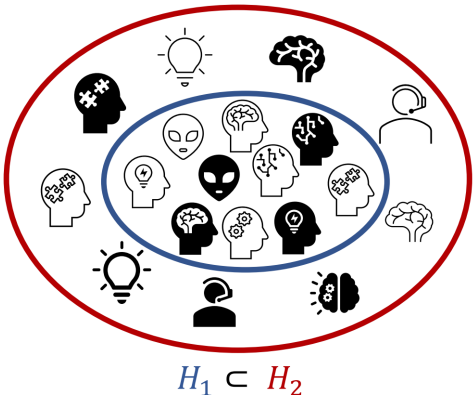

Figure 4: Trade Offs

Here H_2 is richer which can make it contain better models as well as harder to learn. We can characterize the sample complexity of learning H using Binary classification, with Binary classification with adversarial perturbations or with Density estimation.

In the case of binary classification the VC-dimension quantifies complexity:

$$m(\epsilon) = \Theta\left(\frac{VC(H)}{\epsilon^2}\right)$$

The upper bound here is achieved using simple ERM

$$L(S,h) = \frac{1}{|S|}\sum_{(x,y)\in S} l(h,x,y)$$

$$h = argmin_{h\in H} L(S,h)$$

And then for uniform convergence:

$$sup_{h\in H}|L(S,h) - L(D_z,h)| = O\left(\sqrt{\frac{VC(H)}{|S|}}\right) w.p. > 0.99 \tag{19}$$

We now introduce sample compression as an alternative.

3.5 Sample Compression

The idea is to try and answer how should we go about compressing a given training set? In classic information theory, we would compress it into a few bits. In the case of sample compression, we want to try to compress it into a few samples.

If we just take the simple example of linear classification Number of required bits is unbounded (depends on the sample).

It has already been shown by Littlestone and Warmuth [10] that Compressibility $\implies$ Learnability

$$m(\epsilon) = \tilde{O}\left(\frac{k}{\epsilon^2}\right)$$

It has also been shown by Moran and Yehudayoff [11] Compressibility $\impliedby$ Learnability

$$k = 2^{O(VC)}$$

Conjecture 1 (Compression Conjecture)**.** $k = \Theta(VC)$

Sample Compression can be very helpful by:

- being simpler and more intuitive
- being more generic. It can work even if uniform convergence fails! Can show optimal SVM bound and we can also perform compression for learning under Adversarial Perturbations.

Typical classifiers are often:

- Sensitive to "adversarial" perturbations, even when the noise is "imperceptible"
- Vulnerable to malicious attacks
- Ignore the "invariance" or domain-knowledge

In the case of classification with adversarial perturbations we had $l^{0/1}(h, x, y) = 1\{h(x) \neq y\}$ and $l^U(h, x, y) = sup_{\bar{x} \in U(x)} l^{0/1}(h, \bar{x}, y)$

and then we will have the L^U be $L^U(D_Z, h) = E_{(x,y)\sim D_Z} l^U(h, x, y)$.

Now, $A_{Z,H}$ probably approximately correct learns H with $m(\epsilon)$ samples if for all D_Z and for all $\epsilon \in (0, 1)$. Now if $S \sim D_Z^{m(\epsilon)}$ then:

$$\Pr_S[L^U(D_Z, A_{Z,H}(S))\epsilon + OPT] < 0.01 \tag{20}$$

However one of the problems with this is if the robust ERM works for all H

$$L(S, h) = \frac{1}{|S|} \sum_{(x,y)\in S} l(h, x, y)$$

$$h = argmin_{h\in H} L(S, h)$$

The robust ERM would not work for all H, uniform convergence can fail,

$$sup_{h\in H}|L^U(S,h) - L^U(D_z,h)|$$

can be unbounded.

We can say that any "proper learner" (outputs from H) can fail.

In a compression-based method the decoder should recover the labels of the training set and their neighbors and then compress the inflates set:

$$k = 2^{O(VC)}$$

So,

$$m^U(\epsilon) = O\left(\frac{2^{VC(H)}}{\epsilon^2}\right) \tag{21}$$

There is an exponential dependence on $VC(H)$.

Ashtiani et al. [9] introduced tolerant adversarial learning $A_{Z,H}$ PAC learns H with $m(\epsilon)$ samples

if $\forall D_Z$, $\forall \epsilon \in (0,1)$, if $S \simeq D_Z^{m(\epsilon)}$ then

$$Pr_S[L^U(D_Z, A_{Z,H}(S)) > \epsilon + inf_{h\in H} L^V(D_Z, A_{Z,H}(S))] < 0.01 \tag{22}$$

And,

$$m^{U,V}(\epsilon) = \tilde{O}\left(\frac{VC(H)d\log(1+\frac{1}{\gamma})}{\epsilon^2}\right) \tag{23}$$

The trick is to avoid compressing an infinite set and now our new goal is that the decoder should only recover labels of things in $U(x)$.

To do so we can define a noisy empirical distribution (using $V(x)$) and then use boosting to achieve a super small error with respect to this distribution. And then, we encode the classifier using the samples used to train weak learners and the decoder smooths out the hypotheses.

It is interesting to think of Why do we need tolerance? There do exist some other ways to relax the problem and avoid $2^{O(VC)}$

- bounded adversary
- Limited black-box query access to the hypothesis
- Related to the certification problem

This is also observable in the density estimation example.

3.6 Gaussian Mixture Models

Gaussian mixture Models are very popular in practice and are one of the most basic universal density approximators. These are also the building blocks for more sophisticated density classes and can think of them as multi-modal versions of Gaussians.

$$f(x) = w_1 N(x|\mu_1, \sum 1) + w_2 N(x|\mu_2, \sum 2) + w_3 N(x|\mu_3, \sum 3)$$

We say F is Gaussian Mixture Model with k components in $\mathbb{R}^d$. And we want to ask how many samples is needed to recover $f \in F$ within L_1 error ϵ.

The number of samples $\simeq m(d, k, \epsilon)$.

To learn single Gaussian in $\mathbb{R}^d$ then

$$O\left(\frac{d^2}{\epsilon^2}\right) = O\left(\frac{\#params}{\epsilon^2}\right)$$

samples are sufficient (and necessary).

Now if we have k Gaussian in $\mathbb{R}^d$ then we want to know if

$$O\left(\frac{kd^2}{\epsilon^2}\right) = O\left(\frac{\#params}{\epsilon^2}\right)$$

samples are sufficient?

There have been some results on learning Gaussian Mixture Models.

Let us take the example of this graph. For a moment look at this as a binary classification problem. The decision boundary has a simple quadratic form!

$$VC - dim = O(D^2)$$

Here "Sample compression" does npt make sense as there are no "labels".

3.7 Compression Framework

We have F which is a class of distributions (e.g. Gaussians) and we have. If A sends t points from m points and B approximates D then we say F admits (t, m)-compression.

Theorem 7. *If F has a compression scheme of size (t, m) then sample complexity of learning F is*

$$\tilde{O}\left(\frac{t}{\epsilon^2} + m\right)$$

$\tilde{O}(\cdot)$ hides polylog factors.

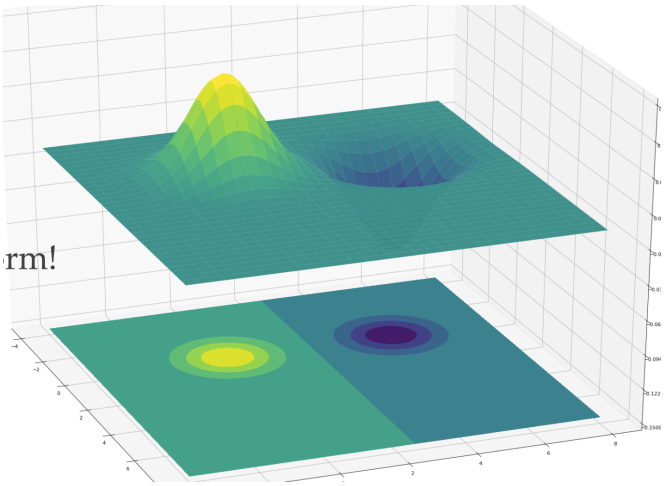

Figure 5: A sample problem

Small compression schemes imply sample-efficient algorithms.

Theorem 8. *If F has a compression scheme of size (t, m) then k mixtures of F admits (kt, km) compression.*

Distribution compression schemes extend to mixture classes automatically! So for the case of GMMs in $\mathbb{R}^d$ it is enough to come up with a good compression scheme for a single Gaussian!

For learning mixtures of Gaussians, the encoding center and axes of ellipsoid is sufficient to recover $N(\mu, \Sigma)$. This admits $\tilde{O}(d^2, \frac{1}{\epsilon})$ compression! The technical challenge is encoding the d eigenvectors "accurately" using only d^2 points.

$\frac{\sigma_{max}}{\sigma_{min}}$ can be large which is a technical challenge.

3.8 Conclusion

- Compression is simple, intuitive, generic
- Compression relies heavily on a few points
  - But still can give "robust" methods
  - Agnostic sample compression
  - Robust target compression
- Target compression is quite general
  - Reduces the problem to learning from finite classes
  - Does it characterize learning?

References

[1] Omar Alrabiah, Venkatesan Guruswami, Pravesh Kothari, and Peter Manohar. A near-cubic lower bound for 3-query locally decodable codes from semirandom CSP refutation. Technical Report TR22-101, Electronic Colloquium on Computational Complexity (ECCC), July 2022.

[2] Arnab Bhattacharyya, L. Sunil Chandran, and Suprovat Ghoshal. Combinatorial lower bounds for 3-query ldcs, 2019. URL https://arxiv.org/abs/1911.10698.

[3] Venkatesan Guruswami, Pravesh K. Kothari, and Peter Manohar. Algorithms and certificates for boolean csp refutation: "smoothed is no harder than random", 2021. URL https://arxiv.org/abs/2109.04415.

[4] Alexander S. Wein, Ahmed El Alaoui, and Cristopher Moore. The kikuchi hierarchy and tensor pca, 2019. URL https://arxiv.org/abs/1904.03858.

[5] Charlie Carlson, Ewan Davies, Nicolas Fraiman, Alexandra Kolla, Aditya Potukuchi, and Corrine Yap. Algorithms for the ferromagnetic potts model on expanders, 2022. URL https://arxiv.org/abs/2204.01923.

[6] Matthew Coulson, Ewan Davies, Alexandra Kolla, Viresh Patel, and Guus Regts. Statistical Physics Approaches to Unique Games. In Shubhangi Saraf, editor, *35th Computational Complexity Conference (CCC 2020)*, volume 169 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 13:1–13:27, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. ISBN 978-3-95977-156-6. doi: 10.4230/LIPIcs.CCC.2020.13. URL https://drops.dagstuhl.de/opus/volltexte/2020/12565.

[7] Peter Frankl and Zoltán Füredi. A short proof for a theorem of harper about hamming-spheres. *Discrete Mathematics*, 34(3):311–313, 1981.

[8] Tyler Helmuth, Will Perkins, and Guus Regts. Algorithmic pirogov-sinai theory. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, page 1009–1020, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367059. doi: 10.1145/3313276.3316305. URL https://doi.org/10.1145/3313276.3316305.

[9] Hassan Ashtiani, Vinayak Pathak, and Ruth Urner. Adversarially robust learning with tolerance, 2022. URL https://arxiv.org/abs/2203.00849.

[10] Nick Littlestone and Manfred Warmuth. Relating data compression and learnability. 1986.

[11] Shay Moran and Amir Yehudayoff. Sample compression schemes for vc classes. *Journal of the ACM (JACM)*, 63(3):1–10, 2016.