Informational-computational trade-off in randomized numerical linear algebra

Xiaobo Yang

April 2023

Contents

Preface

Explore information-theoretic lower bound and computational threshold(an upper bound yielded by algorithm) in randomized numerical linear algebra. We hope the gap is as small as possible.

Chapter 1

Practical Applications

Some coding and practical projects.

1.1 Datasets

xiaobo: wiki,arxiv matrix [Meyer et al., 2021]

1.2 Statistical applications

1.2.1 Gaussian process

xiaobo: train gaussian process in [Jiang et al., 2021] [Meyer et al., 2021]

Chapter 2

Theory for application

Specific complexity theory of specific applications in some areas.

2.1 Matrix-vector multiplication queries

2.1.1 Low rank approximation

[Bakshi et al., 2022]

xiaobo: reduced to estimation of minimum eigenvalue [Braverman et al., 2021]

xiaobo: lra is basically power method, which must be adaptive, is there possibility to find non-adaptive algorithm to estimate lra, or simply estimate first k eigenvalue(k is small)

2.1.2 Solve linear system

[Braverman et al., 2021]

xiaobo: reduced to estimation of minimum eigenvalue [Braverman et al., 2021]

Chapter 3

Theory for areas

General complexity theory of specific areas of problems.

3.1 Matrix-vector multiplication queries

We interact with the algorithm only through inputting query vectors and outputting matrix-vector product oracles.

3.1.1 Trace estimation

Informational threshold

Problem 3.1.1 (trace estimation [Meyer et al., 2021]). *Given a matrix A, we input query vectors $r_1, ..., r_m$ and output product $Ar_1, ...Ar_m$ to estimate $tr(A)$.*

There are a two types of query vectors.

1. (Adaptive) Vectors $r_1, ..., r_m$ are chosen adaptively. For example, $r_2 = Ar_1$. In this case, A may be used for several times.
2. (Non-adaptive) Vectors $r_1, ..., r_m$ are chosen independently. In this case, A can be used for only once, for example, streaming data.

Theorem 3.1.1 (Lower bound of adaptive queries [Meyer et al., 2021]). *Let A be a PSD matrix. If $r_1, ..., r_m$ are adaptive and with integer entries in $\{-2^b, ..., 2^b\}$, then we need at least $m = \Omega\left(\frac{1}{\epsilon(b+\log(1/\epsilon))}\right)$ queries to output an estimate t so that, with probability at least $2/3$, $(1-\epsilon)tr(A) \le t \le (1+\epsilon)tr(A)$.*

Proof sketch for Theorem 3.1.1. First, reduce the problem to Gap-hamming distance problem(GHD) 4.1.1. For any vector $s, t \in \mathbb{R}^n$, reshape s, t to matrix $S, T \in \mathbb{R}^{n\times n}$. Let $Z = S + T$, $A = Z^T Z$.

Then, if we estimate $tr(A)$ with error $\epsilon = 1/\sqrt{n}$, we have

$$\widehat{tr(A)} \ge 2(n+\sqrt{n})(1-1/\sqrt{n}) = 2n-2 \quad if \quad \langle s,t\rangle \ge \sqrt{n},$$
$$\widehat{tr(A)} \le 2(n-\sqrt{n})(1+1/\sqrt{n}) = 2n-2 \quad if \quad \langle s,t\rangle \le -\sqrt{n},$$

which means we can solve GHD problem by simply compare $\widehat{tr(A)}$ with $2n-2$.

Finally, use lower bound of GHD. The reduction above cost $O(m\sqrt{n}(\log n + b)) = \Omega(n)$ bits by Theorem 4.1.1. Note that $n = 1/\epsilon^2$, thus we have $m = \Omega\left(\frac{1}{\epsilon(b+\log(1/\epsilon))}\right)$. □

Theorem 3.1.2 (Lower bound of non-adaptive queries [Meyer et al., 2021]). *Let A be a PSD matrix. If $r_1, ..., r_m$ are non-adaptive, then we need at least $m = \Omega\left(\frac{1}{\epsilon}\right)$ queries to output an estimate t so that, with probability at least $3/4$, $(1-\epsilon)tr(A) \le t \le (1+\epsilon)tr(A)$.*

Proof sketch for Theorem 3.1.2. First, solving the problem implicitly solves a testing problem with m matrix-vector product oracles.

Problem 3.1.2 ([Meyer et al., 2021]). *Fix $d, n \in \mathbb{N}$ such that $d \geq n$ and $n = 1/\epsilon$ for $\epsilon \in (0, 1]$. Let $A = G^T DG \in \mathbb{R}^{d\times d}$, where $G \in \mathbb{R}^{n\times d}$ generated by i.i.d $\mathcal{N}(0, 1)$ and $D = I_n := D_1$ or $\begin{bmatrix} I_{n-1} & 0 \\ 0 & 0 \end{bmatrix} := D_2$. Consider any algorithm taking a query matrix $U \in \mathbb{R}^{d\times m}$ as input and outputting product oracle $AU \in \mathbb{R}^{d\times m}$ to identify D.*

By concentration inequality 5.2.3 of Hutchinson's estimator, we have that with probability at least $11/12$,

$$\frac{1}{d}tr(G^T D_1 G) \geq (1 - \epsilon/4)tr(D_1) \quad and \quad \frac{1}{d}tr(G^T D_2 G) \leq (1 + \epsilon/4)tr(D_2).$$

Then, if we estimate $t \approx tr(A)$ with probability at least $3/4$ and error $\epsilon/4$, we have

$$t \geq (1 - \epsilon/4)tr(A) \geq (1 - \epsilon/4)^2 d/\epsilon \geq (1 - \epsilon/2)d/\epsilon \quad if\ D = D_1,$$
$$t \leq (1 + \epsilon/4)tr(A) \leq (1 + \epsilon/4)^2(1 - \epsilon)d/\epsilon \leq (1 - \epsilon/2)d/\epsilon \quad if\ D = D_2,$$

which means we can identify D by simply compare t with $(1 - \epsilon/2)d/\epsilon$. By union bound, the probability is at least $2/3$.

Then, this problem can be further reduced to problem 4.1.5. Let $M = \begin{bmatrix} N^T N \\ L^T N \end{bmatrix}$, where $N \perp\!\!\!\perp L$ and $L \in \mathbb{R}^{n\times(d-m)}$ with $i.i.d\ N(0, 1)$. Then we have $M \overset{\text{dist}}{=} G^T DGE_m$, where $E_m = [e_1, ..., e_m]$ is the first m standard basis vectors. Therefore, it suffices to run algorithm above and if the output $D = D_1$ we claim $C = I$ otherwise $C = I - zz^T$.

Finally, use lower bound theorem 4.1.5, we have $m = \Omega\left(\frac{1}{\epsilon}\right)$. □

Remark 3.1.1.

1. *Lower bound of adaptive case is smaller than non-adaptive case. It make senses since adaptive queries contain more information on A.*
2. *The difference of two bounds is mathematically because intermediate problem 3.1.2 requires oracles(columns of U) are non-adaptive. But GHD problem doesn't have any requirement on relationships between oracle r s.*
3. *This inspires us, to reduce a problem to get a lower bound, we must keep key structures of problems as many as possible.*

   For instance, although we can reduce the non-adaptive scenario to GHD, the resulting bound is significantly weaker. This is because we abandon the critical constraint of exclusively using independent oracles r s to tackle the problem, which renders it more challenging.

Later, [Jiang et al., 2021] improve the lower bound with error probability δ.

Theorem 3.1.3 (Lower bound of non-adaptive queries with δ-error [Jiang et al., 2021]). *Let A be a PSD matrix. If $r_1, ..., r_m$ are non-adaptive, then we need at least $m = \Omega\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon} + \frac{\log(1/\delta)}{\log\log(1/\delta)}\right)$ queries to output an estimate t so that, with probability at least $1 - \delta$, $(1 - \epsilon)tr(A) \leq t \leq (1 + \epsilon)tr(A)$.*

Proof sketch for Theorem 3.1.3. The proof is divided into two parts according to whether ϵ is small or large.

First, for small $\epsilon = O(1/\sqrt{\log(1/\delta)})$ we have $m = \Omega\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon}\right)$:

1. Use Yao's principle(theorem 5.1.1) we only need to find a hard distribution. Since the bound is required to satisfied by any size n of matrix $A \in \mathbb{R}^{n\times n}$, we consider large n. Use theorem 5.4.1 to quantify the remaining randomness after we take several matrix-vector product oracles, we can observe if the number of queries is not so large, then the remaining randomness will yield large variance and thus large error probability.

2. Specifically, let $W = I_n + \frac{1}{2C\sqrt{n}}(G + G^T)$ where entries of G are *i.i.d.* $\mathcal{N}(0,1)$. Then, W is PSD by theorem 5.4.2 and $tr(W) \leq 2n$ *w.h.p.* for large n, since $tr(G) \sim \mathcal{N}(0,n)$. Quantitatively, for any $\delta \in (0,1)$, take $n = \Omega(\log(1/\delta))$ then $tr(W) \leq 2n$ with probability at least $1 - \delta/10$. Then, for any $\epsilon = O(\frac{\sqrt{\log(1/\delta)}}{n}) = O(1/\sqrt{\log(1/\delta)})$, if we just take $m = n/2 = O\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon}\right)$, use theorem 5.4.1 we prove the best estimation precision any algorithm can achieve *w.h.p.* is at most $c\sqrt{\log(1/\delta)}$ for $c = \Omega(1)$, which is large than required precision ϵ and thus yield lower bound $m = \Omega\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon}\right)$.

Then, for any $\epsilon \in (0,1)$, we have uniformly lower bound $m = \Omega\left(\frac{\log(1/\delta)}{\log\log(1/\delta)}\right)$:

1. The problem is further reduced to a testing problem 4.1.6. The reduction is naturally to consider trace estimation of a random matrix B generated from P or Q and find a gap between both confidence intervals. Actually, by concentration theorem 5.4.2 we have $W + 6\sqrt{\log(1/\delta)}I_n$ is PSD *w.h.p.*, and $|tr(W)| \leq 2\sqrt{2}\log(1/\delta)$ since $tr(W) \sim \mathcal{N}(0, 4\log(1/\delta))$. Then, for any non-adaptive trace estimation algorithm with precision ϵ, consider constant $C > \frac{10(1+\epsilon)}{1-\epsilon} - 6$, then for trace estimation t, we have

   - if $B \sim P$, then $t \geq (1-\epsilon)tr(B) = (1-\epsilon)((C+6)\log^{3/2}(1/\delta) - 2\sqrt{2}\log(1/\delta)) > 6(1+\epsilon)\log^{3/2}(1/\delta)$ with probability at least $1 - 2\delta$,
   - if $B \sim Q$, then $t \leq (1+\epsilon)tr(B) \leq (1+\epsilon)(6\log^{3/2}(1/\delta) + 2\sqrt{2}\log(1/\delta)) < 6(1+\epsilon)\log^{3/2}(1/\delta)$ with probability at least $1 - 2\delta$,

   which distinguishes P and Q.

2. Use hardness theorem 4.1.6 we require at least $m = \Omega\left(\frac{\log(1/\delta)}{\log\log(1/\delta)}\right)$ queries.

Finally, combining these two bound by simple summation, we have our final lower bound. □

Remark 3.1.2. *Add spiked component $C\log^{3/2}(1/\delta)$ in the testing problem may be a standardized trick to analyze information-theoretic lower bound. Tuning the magnitude of the spiked component, we can show the phase transition by information-theoretic toolkit.*

Later, [Woodruff et al., 2022] shows lower bound for general schatten p norm in both bit complexity(consider total complexity of communication of bits) and RAM model(only consider number of queries).

Theorem 3.1.4 (Lower bound of adaptive queries, general matrices and schatten p norm [Woodruff et al., 2022])**.** *Let A be a square matrix. If $r_1, ..., r_m$ are adaptive, then we need at least,*

- *(bit complexity) for queries with entries specified by b bits(integers range in $[-2^b, 2^b]$),*
  - *($p \in [1,2]$) $m = \Omega\left(\frac{1}{\epsilon^p(b+\log(1/\epsilon))} + \frac{\log(1/\delta)}{b+\log\log(1/\delta)}\right)$*
  - *(any p) $m = \Omega\left(\frac{\log(1/\delta)}{b+\log\log(1/\delta)}\right)$*
- *(RAM model) for $\epsilon \in (0, \log^{1/2-1/p}(1/\delta))$,*
  - *($p \in [1,2]$) $m = \Omega\left(\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon}\right)^p\right)$*

queries to output an estimate t so that, with probability at least $1-\delta$, $|t - tr(A)| \leq \epsilon\|A\|_p$.

Proof sketch for Theorem 3.1.4. For bit complexity, the proof is based on communication problem of approximate orthogonality and gap inequality

- ($p \in [1,2]$)
  1. First, we show any algrotihm $\mathcal{A}$ solve trace estimation of a matrix with size n, with error $\epsilon\|\cdot\|_p$ and probability at least $1-\delta$, can explicitly solve a ORT_{b,n^2} problem 4.1.2. To do that, assume Alice has $a \in \{\pm 1\}^{n^2}$ and Bob has $b \in \{\pm 1\}^{n^2}$. Reshape the vector into two matrices $A, B \in \mathbb{R}^{n\times n}$ such that $\langle a, b\rangle = tr(AB)$. Now the task is to compare $|tr(AB)|$ with bn.

In each round, we give a query q_i and compute q_iA and send it to Bob, then Bob compute q_iAB and send it back to alice. Assume that after $r(n)$ rounds, we got a good trace estimation t through $q_1AB, ..., q_{r(n)}AB$. Formally, we have

$$\mathbb{P}_{\mathcal{A}}(|t - tr(AB)| \geq \epsilon \|AB\|_p | A, B) \leq \delta.$$

2. Then, assume $a, b \sim Unif\{\{\pm 1\}^{n^2}\}$ independently, by Holder's inequality for $p \in [1, 2]$ and Markovian's inequality we have $\mathbb{P}_{A,B}(\epsilon\|AB\|_p \geq \epsilon\sqrt{t}n^{1/p+1}) \leq \mathbb{P}_{A,B}(\epsilon\|AB\|_F \geq \epsilon\sqrt{t}n^{3/2}) \leq 1/t, \forall t > 0$. Therefore, take δ', t such that $1/\delta' = t = \left(\frac{1}{\epsilon n^{1/p}} \frac{b}{10}\right)^2$, we have $\mathbb{P}_{A,B}(\epsilon\|AB\|_p \geq \frac{bn}{10}) \leq \delta'$ and thus

$$\mathbb{P}_{\mathcal{A},A,B}\left(|t - \langle a, b\rangle| \geq \frac{bn}{10}\right) \leq \delta + \delta'.$$

By this, we can solve ORT_{b,n^2} *w.h.p.* by comparing $|t|$ with bn.

3. Finally, it's easy to check bit complexity each round is $O(\log(2^b n)) = O(b + \log n)$. Therefore, by theorem 4.1.2, we have $r(n) \cdot O(n(b + \log n)) = \Omega(n^2)$, thus $r(n) = \Omega\left(\frac{n}{b+\log n}\right)$. Since $\delta' = O(1)$, the analysis above only works for $n = O(1/\epsilon^p)$, thus the best lower bound we can achieve by this method is $r(1/\epsilon^p) = \Omega\left(\frac{1}{\epsilon^p(b+\log(1/\epsilon))}\right)$. Combine this and the results below we obtain the final bound.

- (any p)

1. First, note that any algrotihm $\mathcal{A}$ solve trace estimation of a matrix with size n, with error $\epsilon\|\cdot\|_p$ and probability at least $1 - \delta$ can explicitly solve a problem 4.1.3 EQ_n. To do that, assume Alice has $a \in \{0, 1\}^n$ and Bob has $b \in \{0, 1\}^n$. Consider a matrix $A = (a - b)(a - b)^T$ and we use $\mathcal{A}$ to estimate $tr(A)$. Then no matter what ϵ is the task is to detect whether $t = 0$. Assume the algorithm need at most $r(n)$ queries to solve that.

2. Then, note that if $a = b$, then for any sequence of queries $q_1, q_2, ...q_{r(n)}$, the oracles Aq_i will always return 0. But once there is one query q such that $Aq \neq 0$, we can stop the algorithm. Thus there is no difference between adaptive and non-adaptive queries(we don't need adaptive any more). Thus it suffices to consider a matrix Q with (non-adaptive) queries as its columns. It suffices to compute whether $Qx = Qy$. So it suffices to let Alice send Qx to Bob and the bit complexity is $O(r(n)(b + \log n))$.

3. Finally, by theorem 4.1.3, we have $r(n) = \Omega\left(\frac{n}{b+\log n}\right)$. In order to solve EQ_n for all possible a, b simultaneously, we require $n = O(\log(1/\delta)$ by union bound. Thus the best lower bound we can achive by this method is $\Omega\left(\frac{\log(1/\delta)}{b+\log\log(1/\delta)}\right)$.

   xiaobo: not so rigorously here, since this method is a random protocol to solve EQ_n, but theorem 4.1.3 is for deterministic protocol

For RAM model, the proof is based on remainning randomness of conditional matrix-vector products(theorem 5.4.1)

1. First, by Yao's principle(theorem 5.1.1), it suffices to construct a hard distribution. Here we take random matrix $W = (G + G^T)/2 \in \mathbb{R}^{n\times n}$ such that G is *i.i.d.* standard gaussian matrix. By theorem 5.4.2 and Holder's inequality , $\|W\|_p \lesssim n^{1/2+1/p}$ *w.h.p.* for $p \in [1, 2]$ and $n = \Omega(\log(1/\delta))$. Therefore, for any algorithm $\mathcal{A}$ that can estimate the trace, using m queries to output an estimation t, with error ϵ and failure probability δ, we have

$$1 - \delta \leq \mathbb{P}_{\mathcal{A},W}(|t - tr(W)| \leq \epsilon\|W\|_p) \leq \mathbb{P}_{\mathcal{A},W}(|t - tr(W)| \lesssim \epsilon n^{1/2+1/p}).$$

2. Then, assume queries we need to estimate $tr(W)$ is $m < n/2$, then by theorem 5.4.1 there is a submatrix $\widetilde{W}$ whose trace $tr(\widetilde{W})$ has variance of $n - m = \Omega(n)$ even conditioned on all matrix-vector queries and oracles. Thus, such an algorithm $\mathcal{A}$ must satisfy

$$\mathbb{P}_{\mathcal{A},W}(|t - tr(W)| \gtrsim \sqrt{\log(1/\delta)}n^{1/2}) = \Omega(\delta).$$

3. Finally, combine both inequalities above, we need $\epsilon n^{1/2+1/p} = \Omega(\sqrt{\log(1/\delta)} n^{1/2})$, which yields $n = \Omega\left(\left(\frac{\sqrt{\log(1/\delta}}{\epsilon}\right)^p\right)$.

□

Theorem 3.1.5 (Lower bound of non-adaptive queries, general matrices and schatten p norm [Woodruff et al., 2022]). *Let A be a square matrix. If $r_1, ..., r_m$ are adaptive, then we need at least,*

- *(bit complexity) for queries with entries specified by b bits(integers range in $[-2^b, 2^b]$),*
  - *(any p) $m = \Omega\left(\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon}\right)^p \frac{1}{b+\log(1/\epsilon)}\right)$*
- *(RAM model) for $\epsilon \in (0, \log^{1/2-1/p}(1/\delta))$,*
  - *($p \in [1,2]$) $m = \Omega\left(\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon}\right)^p + \frac{\log(1/\delta)}{\log\log(1/\delta)}\right)$*

queries to output an estimate t so that, with probability at least $1-\delta$, $|t - tr(A)| \leq \epsilon \|A\|_p$.

Proof sketch for Theorem 3.1.5. For bit complexity, the proof is based on communication problem of augmented indexing

xiaobo: However, the original proof in [Woodruff et al., 2022] is in a mess, maybe it's wrong, need to check [JW13]

1. First,

For RAM model, the proof is similar to [Jiang et al., 2021]

1. First,

□

A more genral version of trace estimation problem is its dynamic extension.

Problem 3.1.3 (Dynamic trace estimation [Woodruff et al., 2022]). *Let $A_1, A_2, ..., A_m$ be square matrices such that*
- *$\|A_i\|_p \leq 1$ for all i.*
- *$\|A_{i+1} - A_i\|_p \leq \alpha < 1$ for all $i \leq m-1$.*

For any $\epsilon, \delta > 0$, estimate $tr(A_i)$ by t_i with $\mathbb{P}(|t_i - tr(A_i)| > \epsilon) \leq \delta$.

Corollary 3.1.1 (xiaobo: general case(adaptive)Lower bound of dynamic trace estimation [Woodruff et al., 2022]). *To solve problem 3.1.3, we need at least*

- *(bit complexity) for queries with entries specified by b bits(integers range in $[-2^b, 2^b]$),*
- - *($p \in [1,2]$) $m = \Omega\left(\alpha m \left(\frac{1}{\epsilon^p(b+\log(1/\epsilon))} + \frac{\log(1/\delta)}{b+\log\log(1/\delta)}\right)\right)$*
  - *($p \in [1,2]$, $\epsilon, \delta \in (0, 1/4)$) $m = \Omega\left(m \min\left(1, \frac{\alpha}{\epsilon}\right) \frac{\log(1/\delta)}{b+\log\log(1/\delta)}\right)$*
- *(RAM model) for $\epsilon \in (0, \log^{1/2-1/p}(1/\delta))$,*
  - *($p \in [1,2]$) $m = \Omega\left(\alpha m \left(\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon}\right)^p\right)\right)$*

queries of matrix-vector product oracles.

Proof sketch for corollary 3.1.1. It's just a static-to-dynamic reduction [Dharangutte and Musco, 2021] combining adaptive case static trace estimation(theorem 3.1.4)

Lemma 3.1.1 (static-to-dynamic reduction [Dharangutte and Musco, 2021]).

□

Computational threshold

In this problem, the **gap between informational and computational threshold** is small(nearly matching), roughly an $O(\log\log(1/\delta))$ constant. The intuition is to improve Hutchinston's estiator by sketching a low rank approximation of A and estimating the remaining component by Hutchinson's estiator, rather than estimate A itself. This induces a trade-off between sketching precision and Hutchinson's estimation, optimizing the trade-off can reduce the variance.

Algorithm 1 Hutch++: randomized trace estimation with **adaptive** matrix-vector queries

Input: Matrix-vector product oracle for PSD matrix $A \in \mathbb{R}^{n\times n}$. Number m of queries.
Output: Approximation to $tr(A)$.
1: Sample $S \in \mathbb{R}^{n\times m/3}, G \in \mathbb{R}^{n\times m/3}$ with *i.i.d.* $\mathcal{N}(0,1)$ entries.
2: Compute an orthonormal basis $Q \in \mathbb{R}^{n\times m/3}$ for the span of AS via QR decomposition.
3: **return** $t = tr(AQQ^T) + \frac{3}{m}tr(G^TA(I-QQ^T)G)$.

Remark 3.1.3. *Here, we take the randomness of Gaussian [Jiang et al., 2021], while the original verision is rademacher's ± 1 variable [Meyer et al., 2021]. Note that, Hutchinson's analysis works for general mean zero sub-Gaussian random variables.*

Theorem 3.1.6 (Upper bound of adaptive queries [Meyer et al., 2021])**.** *Let $m = O\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon} + \log(1/\delta)\right)$. Then with probability at least $1-\delta$, Hutch++ yields an estimation t such that $(1-\epsilon)tr(A) \le t \le (1+\epsilon)tr(A)$.*

Proof sketch for Theorem 3.1.6. First, note that $t = tr(AQQ^T) + H_{m/3}(A(I-QQ^T))$, where $H_l(\cdot)$ is Hutchinson's Estimator(definition 5.2.1). Denote $\Delta := A(I-QQ^T)$ and $\tilde{A} = AQQ^T$. Thus, if $m = O(\log(1/\delta))$, by theorem 5.2.3 we have $|t - tr(A)| = |tr(\Delta) - H_{m/3}(\Delta)| \lesssim \sqrt{\frac{\log(1/\delta)}{m/3}}\|\Delta\|_F$ with probability(of G) at least $1-\delta$.

Then, if $m = O(k + \log(1/\delta))$, by theroem 5.2.5 we have, $AS/\sqrt{m}$ is a $(1/9, 0, k)$-projection-cost-preserving of A(definition 5.2.2) with probability(of S) at least $1-\delta$, and by theorem 5.2.2, since $\|AS - ASQ_kQ_k^T\| = \|AS - (AS)_k\|_F$, we have $\|\Delta\|_F \le \|A - AQ_kQ_k^T\|_F \le \sqrt{\frac{1+1/9}{1-1/9}}\|A - A_k\|_F \le \sqrt{2}\|A - A_k\|_F$ and thus $\|\Delta\|_F \le \sqrt{2}\|A - A_k\|_F$ with probability(of S) at least $1-\delta$.

Therefore, if $k = \frac{\sqrt{\log(1/\delta)}}{\epsilon}$ and $m = O\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon} + \log(1/\delta)\right)$, use lemma 5.5.1, we have $|t - tr(A)| \lesssim \sqrt{\frac{\log(1/\delta)}{mk}}tr(A) \lesssim \epsilon \cdot tr(A)$ *w.h.p.*. □

Algorithm 2 NA-Hutch++: randomized trace estimation with **non-adaptive** matrix-vector queries

Input: Matrix-vector product oracle for PSD matrix $A \in \mathbb{R}^{n\times n}$. Number m of queries.
Output: Approximation to $tr(A)$.
1: Fix constants c_1, c_2, c_3 such that $c_1 > c_2$ and $c_1 + c_2 + c_3 = 1$.
2: Sample $S \in \mathbb{R}^{n\times c_1 m}, R \in \mathbb{R}^{n\times c_2 m}, G \in \mathbb{R}^{n\times c_3 m}$ with *i.i.d.* $\mathcal{N}(0,1)$ entries.
3: **return** $t = tr(AR(S^TAR)^+S^TA) + \frac{1}{c_3 m}tr(G^T(A - [AR(S^TAR)^+S^TA])G)$[1].

xiaobo: the proof seems no requirement of $c_1 < c_2$ as original paper, but $c_2 < c_1$ instead, write **code** to check.(put experiment part in chapter 1

Theorem 3.1.7 (Upper bound of non-adaptive queries [Jiang et al., 2021])**.** *Let $m = O\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon} + \log(1/\delta)\right)$. Then with probability at least $1-\delta$, NA-Hutch++ yields an estimation t such that $(1-\epsilon)tr(A) \le t \le (1+\epsilon)tr(A)$.*

[1] Here M^+ means Moore-Penrose pseudoinverse of M

Proof sketch for Theorem 3.1.7. Similar to the adaptive one, here we divide A into $\tilde{A} := AR(S^TAR)^+S^TA$ and $\Delta := A - \tilde{A}$. Based on the proof of theorem 3.1.6, it suffices to show $\|\Delta\|_F = O(1)\|A - A_k\|_F$ *w.h.p.* and then we can obtain exact the same bound as before. Formally, we will prove if $m = O(k + \log(1/\delta))$, then with probability at least $1 - \delta$ we have $\|\Delta\|_F = O(1)\|A - A_k\|_F$.

Basic idea is to use theorem 5.2.7 to do column and row sketching for regression respectively.

First, note that $\|A - A_k\|_F = \min_X \|A - A_kX\|_F$. Consider regression sketching $\bar{X} = \arg\min_X \|R^T(A - A_kX)\|_F = (R^TA_k)^+R^TA$, then if $c_2m = O(k + \log(1/\delta))$, we have $O(1)\|A - A_k\|_F \geq \|A - A_k\bar{X}\|_F$ with probability(of R) at least $1 - \delta$.

Then, note that $\|A - A_k\bar{X}\|_F = \|A - A_k(R^TA_k)^+R^TA\|_F = \|A - AR(A_kR)^+A_k\|_F \geq \|A - ARX^*\|_F$, where $X^* = \arg\min_X \|ARX - A\|_F$. Then another regression sketching yields that, if $c_1m = O(c_2m + \log(1/\delta))$, we have $O(1)\|A - ARX^*\|_F \geq \|A - AR\widetilde{X}\|_F$, where $\widetilde{X} := \arg\min_X \|S^T(ARX - A)\|_F = (S^TAR)^+S^TA$.

Finally, if $m = O(k + \log(1/\delta))$ and $c_1 > c_2$ we have $O(1)\|A - A_k\|_F \geq O(1)\|A - \bar{X}A_k\|_F \geq O(1)\|A - ARX^*\|_F \geq \|A - AR\widetilde{X}\|_F = \|A - \widetilde{A}\|_F$ □

Remark 3.1.4. *xiaobo: compare the improvement with [Meyer et al., 2021] for non-adaptive queries*

xiaobo: even though the bound is tight for Gaussian, what about other randomness, for example rademacher? [Meyer et al., 2021] uses rademacher

xiaobo: add dynamic alg here

Theorem 3.1.8 (Upper bound, general schatten p norm, dynamic case [Woodruff et al., 2022])**.** *For any $\epsilon, delta > 0$, the algroithm xiaobo: (add ref) solves the problem 3.1.3 using at most*

- *($p \in [1, 2]$) $m = O\left((m\alpha + 1)\log^{1+p}(1/\alpha)\left(\frac{\sqrt{\log(1/(\alpha\delta))}}{\epsilon}\right)^p + m\log(1/(\alpha\delta))\right)$ xiaobo: (thm B.2 in [Woodruff et al., 2022])*
- *($p = 1$) $m = O\left((m\alpha + 1)\log^2(1/\alpha)\left(\frac{\sqrt{\log(1/(\alpha\delta))}}{\epsilon}\right) + m\min(1, \alpha/\epsilon)\log(1/(\alpha\delta))\right)$ xiaobo: (thm 3.1 in [Woodruff et al., 2022])*

(adaptive) queries of matrix-vector oracles.

Proof sketch for Theorem 3.1.8. □

Remark 3.1.5.
1. *For $m = 1$(static case), the upper bound match the lower bound up to logarithm of the parameters.*
2. *For $p = 1$, the first condition can be relaxed to $\|A_1\|_* \leq 1$. See theorem B.3 in [Woodruff et al., 2022].*

3.1.2 Eigenvalue estimation

[Braverman et al., 2021]

3.1.3 Matrix maximal entry estimation

Problem 3.1.4 (trace estimation [Meyer et al., 2021])**.** *Given a matrix A, we input query vectors $r_1, ..., r_m$ and output product $Ar_1, ...Ar_m$ to estimate $\max_{i,j} A_{i,j}$.*

Chapter 4

Theoretic Frameworks

General complexity frameworks for general theoretic areas.

4.1 Meta-complexity

Some classical problems that are usually reduced to.

4.1.1 Communication complexity

Gap-Hamming-Distance problem

Problem 4.1.1 (Gap-Hamming [Meyer et al., 2021]). *Let Alice and Bob be communicating parties who hold vectors $s \in \{\pm 1\}^n$ and $t \in \{\pm 1\}^n$, respectively. The Gap-Hamming problem asks Alice and Bob to return*

$$1 \text{ if } \langle s,t\rangle \geq \sqrt{n} \qquad \text{and} \qquad -1 \text{ if } \langle s,t\rangle \leq -\sqrt{n}.$$

Denote it by $GHD_n(x,y)$.

In general, we need to compute the function $f(x,y)$ for $x \in \mathcal{X}, y \in \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are separate sources and we cannot simultaneously obtain both x and y. Instead, we repeatedly communicate information from x to y or from y to x, accumulating intermediate information during this process. Finally, we use this accumulative intermediate information to compute $f(x,y)$. The goal is to minimize the number of times of communication or the total amount of communication bits.

To formalize this communication process, a binary tree structure called a "protocol tree" T(Fig.4.1) is used. Each possible value of $f(x,y)$ corresponds to a leaf node L in T. The communication complexity is measured by the height $H(T)$ of the tree T. Notably, for any node N in T, the set $R(N) := (x,y) : (x,y)$ can reach node N always forms a square, i.e., $R(N) = A \times B$ for some $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$. As a result, the leaf nodes form a rectangular partition of $\mathcal{X} \times \mathcal{Y}$.

By using the minimum number of square partitions M_f of the function $f(x,y)$, we can derive a lower bound on the communication complexity:

$$H(T) \geq \log_2 M_f.$$

Theorem 4.1.1 ([Chakrabarti and Regev, 2012] [Vidick, 2012] [Sherstov, 2012]). *Any randomized protocol solving problem 4.1.1 GHD_n with probability $2/3$ needs at least $\Omega(n)$ bits of communication.*

Proof sketch for Theorem 4.1.1. The original proof is in [Chakrabarti and Regev, 2012] and [Vidick, 2012] make a simpler one. Here we use proof in [Sherstov, 2012], which is simplist among them.

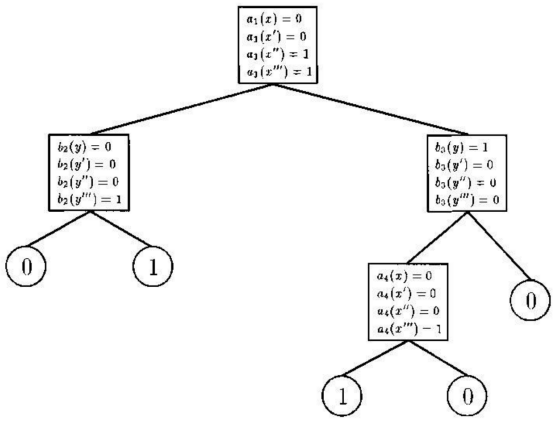


Figure 1.1: A protocol tree

	y	y'	y''	y'''
x	0	0	0	1
x'	0	0	0	1
x''	0	0	0	0
x'''	0	1	1	1

Figure 1.2: The function f computed by the protocol of Figure 1.1

Figure 4.1: Protocol tree [Kushilevitz and Nisan, 1997]

First, use **Yao's principle**(theorem 5.1.1) to reduce the randomness of algorithm to randomness of a distribution μ of input (x, y). Formally, consider all those deterministic protocol tree $T(x, y)$ such that $\mathbb{P}_{(x,y)\sim\mu}[T(x, y) \neq f(x, y)] \leq \epsilon$. Denote the minimum complexity(tree height) of such deterministic T by $R_{\mu,\epsilon}(f)$, and denote minimum expectation complexity of random tree with error ϵ by $C_\epsilon(f)$, we have $C_\epsilon(f) \gtrsim R_{\mu,\epsilon}(f)$. See [Yao, 1983] lemma 2 in sect. 3.3 for details.

Then, it suffices to find a hard distribution μ of (x, y) and lower bound of $R_{\mu,\epsilon}(f)$. This is done by a trick called **ϵ-corruption bound**:

Lemma 4.1.1 (Yao's corruption bound [Sherstov, 2012])**.** *If $\exists\varepsilon, \delta \in (0, 1)$ such that μ satisfies*

$$\mu\left(R \cap f^{-1}(+1)\right) > \varepsilon\mu\left(R \cap f^{-1}(-1)\right)$$

for every rectangle $R \subseteq X \times Y$ with $\mu(R) > \delta$, then $R_{\mu,\xi}(f) \geq \log_2\left(\frac{1}{\delta}\left(\mu\left(f^{-1}(-1)\right) - \frac{\xi}{\varepsilon}\right)\right)$ for small $\xi > 0$.

Intuitively, the misclassification error in any large rectangle is not so large. Therefore, when total error is a given constant, we can lower bound the number of rectangle partition and thus lower bound the tree height. See [Yao, 1983] lemma 3 in sect. 3.3 for originality and details. Usually, people try to directly prove corruption bound for uniformly distribution.

Finally, in our setting, the Gap-Hamming distance problem(GHD_n) is further reduce to Gap-orthogonality problem(ORT_n, problem 4.1.2) with $R_{1/3}(GHD_n) = \Omega(R_{1/3}(ORT_n))$, where $f(x, y) = (-1)\mathbf{1}_{|\langle x,y\rangle|\leq\sqrt{n}} + (+1)\mathbf{1}_{|\langle x,y\rangle|\geq 2\sqrt{n}}$. Note that $f^{-1}(+1) = \{(x, y)||\langle x, y\rangle| \geq 2\sqrt{n}\}$, thus the corruption bound intuitively(not so rigorously) means for rectangle R with large probability, we have

$$\exists c, C > 0, \quad \mathbb{P}_{(x,y)\sim Unif[R]}[|\langle x, y\rangle| > C] \geq c, \tag{4.1}$$

which is essentially a **anti-concentration analysis**. Choosing uniform distribution is enough, it is proved by **probabilistic method** with Talagrand's inequalities(corollary 5.2.1) and a interesting combinatorial accounting trick:

Lemma 4.1.2 (Corruption bound [Sherstov, 2012])**.** *Let μ be uniform distribution on $\{\pm1\}^n \times \{\pm1\}^n$ and $R = A \times B$ be a rectangle such that $\mu(R \cap f^{-1}(+1)) \leq \epsilon\mu(R)$ then $\mu(R) = e^{-\Omega(n)}$. is small. In other word, if $\mu(R)$ is large, we have corruption bound.*

Proof sketch for lemma 4.1.2. The anti-concentration inequality is reduced to that, by corollary 5.2.1 we can find a set I of near-orthogonal vectors of $x \in A$ in rectangle $R = A \times B$ with large probability, and then by some linear algebra and Hoffman-Wielandt inequality, for most of $y \in B$, there must exists a $x' \in I$ such that $|\langle x', y\rangle|$ is large. □

Use the lemma, we compute $C_{1/3}(GHD_n) = \Omega(R_{1/3}(ORT_n)) = \Omega(n)$. □

Remark 4.1.1. *For general communication problem, the analysis of average-case complexity usually follows the roadmap above:*

1. *Use Yao's principle to reduce the problem with randomized algorithm into a randomized input with deterministic algorithm.*
2. *Find the math structures of complexity in setting of deterministic algorithm(In communication problem, it's height of a binary tree whose leaf nodes are rectangle partition).*
3. *Abstract these structures into pure math language, and then use fundamental math tool to work it out(In this communication function $f = GHD_n$, the structure is an anti-concentration analysis and tools are some probabilistic methods).*

Remark 4.1.2. *Another proof [Vidick, 2012] use a different description of the anti-concentration(theorem 5.2.1) in corruption bound (4.1), where they consider general vectors $x, y \in \mathbb{R}^n$ ($\mathcal{X}, \mathcal{Y}, R$ may be continuous).*

Approximate-Orthogonality problem

Problem 4.1.2 ([Chakrabarti et al., 2012] [Woodruff et al., 2022]). *Let Alice and Bob be communicating parties who hold vectors $s \in \{\pm 1\}^n$ and $t \in \{\pm 1\}^n$, respectively. The Approximate-Orthogonality problem asks Alice and Bob to return*

$$1 \ if \ |\langle s, t\rangle| \le b\sqrt{n} \qquad and \qquad -1 \quad otherwise.$$

Denote it by $ORT_{b,n}(x, y)$.

Theorem 4.1.2 ([Chakrabarti et al., 2012] [Woodruff et al., 2022]). *Any randomized protocol solving problem 4.1.2 $ORT_{b,n}$ with probability at least $\Phi(2.01 \max\{66, b\})$ ($b > 1/5$) needs at least $\Omega(n)$ bits of communication. Where $\Phi(x)$ is c.d.f. of $\mathcal{N}(0, 1)$.*

Proof sketch for Theorem 4.1.2. xiaobo: based on anti-concentration lemma 5.2.2 □

Gap-Equality problem

Problem 4.1.3 ([Woodruff et al., 2022]). *Let Alice and Bob be communicating parties who hold vectors $s \in \{0, 1\}^n$ and $t \in \{0, 1\}^n$ with either $s = t$ or $\|s - t\|_2^2 = n/2$, respectively. The Gap-Equality problem asks Alice and Bob to return*

$$1 \ if \ s = t \qquad and \qquad -1 \quad otherwise.$$

Denote it by $EQ_n(x, y)$.

Theorem 4.1.3. *Any deterministic protocol solving problem 4.1.3 EQ_n needs at least $\Omega(n)$ bits of communication.*

Augmented Indexing problem

Problem 4.1.4 ([Woodruff et al., 2022]). *Given set $\mathcal{U}$ and an elements $\perp \notin \mathcal{U}$. Let Alice and Bob be communicating parties who hold vectors $s \in \mathcal{U}^n$ and $t \in \{\mathcal{U} \cup \{\perp\}\}^n$ with either $s_k = t_k,\ k < i$, $t_i \in \mathcal{U}$ and $y_{i+1} = \cdots = y_n = \perp$ for some unique i. The Gap-Equality problem asks Alice and Bob to return*

$$1 \ if \ s_i = t_i \qquad and \qquad -1 \quad otherwise.$$

Denote it by $IND_{n,\mathcal{U}}(x, y)$.

Theorem 4.1.4. *Any one-way(only one people can send message to the other people) randomized protocol solving problem 4.1.4 $IND_{n,\mathcal{U}}(x, y)$ with error $\delta \leq \frac{1}{4|\mathcal{U}|}$ needs at least $n \log |\mathcal{U}|/2$ bits of communication.*

4.1.2 Spiked wishart matrix testing

Problem 4.1.5 (Spiked wishart matrix [Meyer et al., 2021] [Perry et al., 2018]). *Let $n = 1/\epsilon$ and let $z \in \mathbb{R}^n$ be a uniformly random unit vector. Let $N \in \mathbb{R}^{n\times m}$ be m i.i.d Gaussian vector drawn from an n-dimensional $\mathcal{N}(0, C)$, where $C = I$ or $I - zz^T$. Use N to identify C.*

Theorem 4.1.5 ([Meyer et al., 2021]). *To solve problem 4.1.5 with probability at least $2/3$, we need at least $m = \Omega\left(\frac{1}{\epsilon}\right)$.*

Proof sketch for Theorem 4.1.5. First, let P denote the distribution of N under null hypothesis $C = I$, and Q denote the distribution of N under alternative hypothesis $C = I - zz^T$. Any testing statistics ϕ outputs 1 for Q and 0 for P. It suffices to bound the total variation distance $d_{TV}(P, Q)$, since its control the summation of Type-1 and Type-2 errors

$$\min_{\phi} \{P(\phi = 1) + Q(\phi = 0)\} = 1 - d_{TV}(P, Q).$$

Then, by Pinsker's inequality, we have $d_{TV}(Q, P) \leq \sqrt{\frac{1}{2} D_{KL}(Q||P)} \leq \sqrt{\frac{1}{2} D_{\chi^2}(Q||P)}$. Therefore it suffices to upper bound $D_{\chi^2}(Q||P) = \int_N \left(\frac{Q(N)}{P(N)}\right)^2 P(N)dN - 1$. By theorem 5.3.1 with $\beta = -1$ and spiked prior z, we have

$$D_{\chi^2}(Q||P) = \mathbb{E}_{v,v'}\left[(1 - \langle v, v'\rangle^2)^{-m/2}\right] - 1,$$

where $v, v \overset{\text{dist}}{=} z$ are uniformly random unit vectors. Therefore, by classical results of random unit vectors, $p.d.f.$ of $x := \langle v, v'\rangle$ is $p(x) = \frac{\Gamma(n-1)}{2\Gamma((n-1)/2)^2}\left(\frac{1-x^2}{4}\right)^{(n-1)/2-1}$.

Finally, direct calculation yields, if $m = O(1/\epsilon)$, $\mathbb{E}_{v,v'}\left[(1 - \langle v, v'\rangle^2)^{-m/2}\right] < 6/5$ and thus $d_{TV}(Q, P) < 1/3$. Therefore, one of Type-1 and Type-2 errors must be larger than $1/3$. □

Problem 4.1.6 ([Jiang et al., 2021]). *Given $\delta \in (0, 1/2)$, set $n = \log(1/\delta)$. Independently take $g \sim \mathcal{N}(0, I_n)$ and $G \in \mathbb{R}^{n\times n}$ with i.i.d. $\mathcal{N}(0, 1)$ entries. Let $W = G + G^T$. Consider two distributions:*

- *Distribution P: $C \log^{3/2}(1/\delta) \cdot \frac{1}{\|g\|_2^2} gg^T + W + 6\sqrt{\log(1/\delta)} I_n$ for some fixed constant $C > 0$.*
- *Distribution Q: $W + 6\sqrt{\log(1/\delta)} I_n$.*

*Use **non-adaptive** queries $q_1, ...q_m$ and oracles $Aq_1, ..., Aq_m$ to distinguish P and Q. In other word, take a matrix $Q \in \mathbb{R}^{n\times m}$ as query and AQ as oracle.*

Theorem 4.1.6 ([Jiang et al., 2021]). *Any randomized algorithm solving problem 4.1.6 with probability $1-\delta$ need at least $m = \Omega\left(\frac{\log(1/\delta)}{\log\log(1/\delta)}\right)$ queries.*

Proof sketch for Theorem 4.1.6. Similar to the proof of theorem 4.1.5, it is naturally to consider bounding the total variation $d_{TV}(P,Q)$, and by rotational invariance we WLOG assume $Q = E_m$, the first m standard basis vectors.

First, denote by P', Q' the distribution of $BQ(B \sim P$ or Q respectively). Let $L_{P'}, L_{Q'} \in \mathbb{R}^l$ be vectorization of matrices from P', Q'(remove the redundant variable, since B is symmetric), where $l = n + (n-1) + \cdots (n-m+1)$. Observe that conditioned on a realization g we have

$$d_{KL}(P', Q'|g) \le d_{KL}(L_{P'}, L_{Q'}|g) \le \|\mathbb{E}L_{P'} - \mathbb{E}L_{Q'}\|_2^2 = \sum_{i=1}^{m} \left\| C\log^{3/2}(1/\delta)\frac{gg^T}{\|g\|_2^2}e_i \right\|_2^2 = C^2\log^3(1/\delta)\frac{\|g^TQ\|_2^2}{\|g\|_2^2}.$$

The first inequality use data processing inequality(theorem 5.3.3) and the second inequality uses theorem 5.3.2.

Then, find a typical event $\mathcal{E}$ with positive probability. We take $\mathcal{E} = \left\{ \frac{\|g^TQ\|_2^2}{\|g\|_2^2} \le \frac{1}{50C^2n^3} \right\}$ and show that $\mathbb{P}(\mathcal{E}) \ge 10\delta$ if we only take $m = O\left(\frac{\log(1/\delta)}{\log\log(1/\delta)}\right)$ queries. Indeed, assume $m \le n/2$

$$\mathbb{P}(\mathcal{E}) \ge \mathbb{P}(\|g\|_2^2 \ge n/2 | \|g^TQ\|_2^2 \ge 1/(100C^2n^2)) \cdot \mathbb{P}(\|g^TQ\|_2^2 \ge 1/(100C^2n^2)) = \Omega(1)\cdot\Omega((\frac{1}{n\sqrt{m}})^m) = \Omega(e^{-\frac{m}{2}\log(n^2m)}).$$

Therefore, when $m = O\left(\frac{\log(1/\delta)}{\log\log(1/\delta)}\right)$, $\mathbb{P}(\mathcal{E}) \ge 10\delta$.

Finally, conditioned realization $g \in \mathcal{E}$, we have the total variation $d_{TV}(P', Q'|g) \le \sqrt{d_{TV}(P', Q'|g)/2} \le 1/3$ by Pinsker's inequality. This means $\mathbb{P}$[the algorithms make mistake$|g] \ge 1/3$ under P' or Q'. Since $\mathbb{P}(\mathcal{E}) \ge 10\delta$, we have $\mathbb{P}$[the algorithms make mistake$] \ge \delta$ under P or Q if we only take $m = O\left(\frac{\log(1/\delta)}{\log\log(1/\delta)}\right)$ queries. □

Remark 4.1.3. *This bound is independent of error ϵ. Informally, this is because there is a sharp phase transition due to its mathematical structure. Similar structures can be found in proof of theorem 3.1.4, provided by Gap-Equality problem 4.1.3's sharp contrast.*

xiaobo: consider directly bound $d_{TV}(P,Q) \le 1 - 2\delta$ instead?

4.2 Informational-computational gap

The informational-computational gap refers to the phenomenon where certain statistical problems are informationally solvable, yet lack efficient (polynomial-time) algorithms in some regimes.

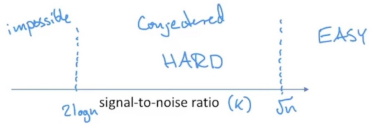
Three major research direction: xiaobo: need to add citations below, see intro of [Bandeira et al., 2022] and appendix A of COLT verision of [Brennan and Bresler, 2020]

1. Reduction form different average-case problems. reduced them to some classical forms, e.g. planted cliques.
2. Prove the failure of the problems under some general classes of algorithms, e.g. sum of squares, low degree polynomial, statistical query model etc.
3. Analyze the phenomena appeared in hard regime, e.g. overlap gap property, low degree likelihood ratio etc.

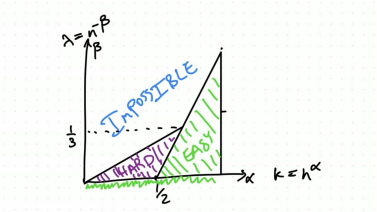
4.2.1 Average-case reduction

Three principles [Brennan and Bresler, 2020]

1. Correctly map the distribution of noise and signal
2. Lower bound from reduction must be tight(achieved by best known efficien algorithm)
3. Tackle the different amount of parameters of the problem to reduce, e.g. biclustering problem has 2 parameters while planted clique has only one.



(a) $PC(n,k)$, picture from [Schramm, 2020]



(b) $BC(n,k,\lambda)$, picture from [Bresler, 2022a]

Figure 4.2: Phase transition of informational and computational thresholds

An example of reduction: reduce biclustering to planted cliques

This section is mainly from [Bresler, 2022a] and [Bresler, 2022b].

Problems

Problem 4.2.1 (Biclustering). *Given a random matrix $X \in \mathbb{R}^{n\times n}$, a biclustering problem with parameter n,k,λ is the following hypothesis testing problem*

- H_0: $X_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$
- H_1: *there exists* $S,T \overset{i.i.d.}{\sim} \binom{[n]}{k}$, $X_{ij} \overset{i.i.d.}{\sim} \begin{cases}\mathcal{N}(\lambda,1) & \text{if } i\in S, j\in T,\\ \mathcal{N}(0,1) & \text{otherwise.}\end{cases}$

Denote it by $BC(n,k,\lambda)$.

Problem 4.2.2 (Planted cliques). *Given a random graph $G=(V,E)$, a planted cliques problem with parameter n,k is the following hypothesis testing problem*

- H_0: $E_{ij} \overset{i.i.d.}{\sim} Ber(1/2)$ *for* $i>j$,
- H_1: *there exists* $U\sim\binom{[n]}{k}$, $E_{ij}=\begin{cases}1 & \text{if } i,j\in U,\\ Ber(1/2) & \text{otherwise.}\end{cases}$

Denote it by $PC(n,k)$.

The phase transition of informational thresholds of both problems can be proved. However, we currently only have evidence for computational thresholds, lacking definitive proof.

- For $PC(n,k)$ with $k=n^\alpha(\alpha>1/2)$, the statistics $T(G)=\mathbf{1}_{\text{max degree}\geq n/2+\sqrt{Cn\log n}}$ for sufficiently large constant C succeeds to detect the hypothesis with polynomial time. This is because
  - Under H_0, $\mathbb{P}(T=1)\leq n\mathbb{P}(Ber(n,1/2)\geq n/2+\sqrt{Cn\log n})\leq n\cdot n^{-2}\to 0$;
  - Under H_1,

$$\begin{aligned}\mathbb{P}(T=0)&\leq \mathbb{P}(\text{for } i \text{ in the planted clique } deg(i)\leq n/2+\sqrt{Cn\log n})\\ &=\mathbb{P}(Ber(n-k,1/2)\leq n/2-k+\sqrt{Cn\log n})^k\leq \mathbb{P}\left(|\sum_{i=1}^{n-k}X_i-\mathbb{E}X_i|\geq k/2+\sqrt{Cn\log n}\right)^k\\ &\leq 2^k e^{-c\frac{k^2}{n}\cdot k}\to 0.\end{aligned}$$

  But there is no efficient algorithm for $\alpha\leq 1/2$.

- For $BC(n,k,\lambda)$ [Ma and Wu, 2015] with $k=n^\alpha$ and $\lambda=n^{-\beta}$,we first show some poly-time statistics in easy regime
  - if $\beta<0$, the maximum statistics $T_{max}(X)=\max_{i,j}X_{i,j}=\begin{cases}O(\sqrt{\log n}) & \text{under } H_0\\ \Omega(\lambda) & \text{under } H_1.\end{cases}$ Thus it succeeds to detect the hypothesis with polynomial time.

– if $\beta > 0$, the average statistics $T_{avg}(X) = \frac{1}{n}\sum_{i,j} X_{i,j} = \begin{cases}\mathcal{N}(0,1) & \text{under } H_0\\ \mathcal{N}(\frac{k^2\lambda}{n},1) & \text{under } H_1.\end{cases}$ Thus it succeeds to detect the hypothesis with polynomial time when $\beta < 2\alpha - 1$.

But there is no efficient algorithm in regime among impossible regime and easy regime, which is conjectured to be hard. Though there does have some statistics solve the problem in this regime, they are not poly-time.

For example, the searching statistics $T_{search} = \frac{1}{k}\max_{|S|=|T|=k}\sum_{i\in S, j\in T} X_{ij} \begin{cases}= O(\sqrt{k\log n}) & \text{under } H_0\\ \geq \frac{k\lambda}{\sqrt{\log n}}\ w.h.p. & \text{under } H_1\end{cases}$, where the first case is because $T_{search} \approx$ max of $\binom{n}{k}^2$ $\mathcal{N}(0,1)$ and the second case is because $T \geq \frac{1}{k}\sum_{i,j:X_{i,j}\sim\mathcal{N}(\lambda,1)} X_{i,j} \sim \mathcal{N}(k\lambda, 1)$.

Therefore $\lambda \gg \frac{\log n}{\sqrt{k}}$ or $\beta < \alpha/2$ can solve the problem. Later we will see this is actually informational boundary.

Theorem 4.2.1. *For $PC(n,k)$, the informational boundary is $2\log n$.*

Proof sketch for Theorem 4.2.1. The proof is based on second moment method, see [Yang et al., 2018] for details. □

Theorem 4.2.2. *For $BC(n,k,\lambda)$, the informational boundary are the line segments shown in figure 4.2b.*

Proof sketch for Theorem 4.2.2. The proof is to prove $d_{TV}(H_0, H_1) \to 0$, see [Brennan et al., 2019] theorem 40 for details. □

Conjecture 4.2.1. *The phase transition of computational thresholds are in figure 4.2.*

Reduction

We will show how to derive computational boundary between easy and hard regime of $BC(n,k,\lambda)$, assuming hardness result of $PC(n,k)$. But first we need the following prelinminary knowledge.

Definition 4.2.1 (Worst-case reduction)**.** *Let P, P' be two worst-case problems. We say that P reduces to P', denoted by $P \leq_p P'$ if there is a polynomial-time computable algorithm $\mathcal{A}$ such that*

$$\mathcal{A}(P) = P',$$

which means operating on the problem instance of P to construct a problem instance of P'. We say that P and P' have equivalent complexity if $P \leq_p P'$ and $P' \leq_p P$.

Example. Reduce 3-SAT to independnent set problem, where the constructed graph G has an independent set of size k iff SAT formula Φ has a satisfying assignment.

Remark. Intuitively, we may think hardness as **"amount of algorithms to solve the problem"**. In that way, $\mathcal{A}(P) = P'$ means all algorithms that can solve P', can also solve P combining with $\mathcal{A}$. Therefore, algorithms for P is no fewer than P' and P is no harder than P'.

Compared to worst-case reduction, average-case reduction is significantly more complex because it involves not only mapping between structurally similar objects, but also mapping between distributions such that their statistical distances are small.

Lemma 4.2.1 (Total variation)**.** *Given two distribution μ and ν, the total variation distance $d_{TV}(\mu,\nu)$ has the following equivalent definitions*

$$\begin{aligned} d_{TV}(\mu,\nu) &= \frac{1}{2}\|\mu-\nu\|_1 = \frac{1}{2}\int |f(x)-g(x)|dx \quad \textit{if they have density } f,g\\ &= \sup_E |\mu(E)-\nu(E)| = \inf_{(X,Y),X\sim\mu,Y\sim\nu} \mathbb{P}(X\neq Y)\\ &= 1 - \min_{\phi:\phi \textit{ is a testing statistics}} \{\mu(\phi \textit{ indicates } \nu) + \nu(\phi \textit{ indicates } \mu)\}. \end{aligned}$$

Lemma 4.2.2. *If random variables X, Y satisfy that $d_{TV}(X,Y) \le \delta$, then there exists a coupling distribution $\mathcal{P}$ such that $\mathbb{P}_{(X,Y)\sim\mathcal{P}}(X = Y) \ge 1 - \delta$.*

Definition 4.2.2 (Reduction)**.** *Let P, P' be two average-case problems. The reduction with error ϵ from P to P' is an polynomial-time computable algorithm $\mathcal{A}$ such that*

$$d_{TV}(\mathcal{A}(P), P') \le \epsilon.$$

Lemma 4.2.3 (Properties of reduction)**.**

- *Let algorithm $\mathcal{A}$ be reduction with error ϵ from P to P'. If algorithm Φ solves P' with probability δ, then $\Phi \circ \mathcal{A}$ solves P with probability $\delta + \epsilon$.*
- *Let algorithm $\mathcal{A}_1$ be reduction with error ϵ_1 from P to P_1, and algorithm $\mathcal{A}_2$ be reduction with error ϵ_2 from P_1 to P_2. Then $\mathcal{A} := \mathcal{A}_2 \circ \mathcal{A}_1$ is reduction with error $\epsilon_1 + \epsilon_2$ from P to P_2.*

Proof for lemma 4.2.3.

1. $\mathbb{P}_{X\sim P}(\phi \circ A \text{ solves } X) \le d_{TV} + \mathbb{P}_{Y\sim P'}(\phi \text{ solves } Y) \le \epsilon + \delta$.
2. $d_{TV}(\mathcal{A}_2 \circ \mathcal{A}_1(P), P_2) \le d_{TV}(\mathcal{A}_2 \circ \mathcal{A}_1(P), \mathcal{A}_2(P_1)) + d_{TV}(\mathcal{A}_2(P_1), P_2) \le \epsilon_1 + \epsilon_2$ by data processing inequality.

□

Now, we try to reduce problem $P = PC(n,k)$ to problem $P' = BC(n,k,\lambda)$. It's natural to map adjacency matrix of the random graph in P to random matrix in P'. However, it's not trivial since the randomness are not the same, and the adjacency matrix is symmetric and without diagonal entries.

To solve these problems, the reduction is constructed by the following steps:

1. Map Bernoulli variables to Gaussian variables.

   To find a mapping Φ such that $\Phi(1) = \mathcal{N}(\lambda, 1)$ and $\Phi(Ber(1/2)) = \mathcal{N}(0,1)$. Denote p.d.f of $\mathcal{N}(\lambda,1), \mathcal{N}(0,1)$ by P, Q respectively. Then p.d.f. of $\Phi(0)$ should be $2Q - P$.

   However, this density may not be positive all the time. Therefore we take truncation p.d.f. such that $\Phi(0) \sim \rho(x) := \frac{1}{Z}(2Q(x) - P(x))\mathbf{1}_{2Q-P>0}$ instead. Though it's not exact mapping we want, but fortunately we have $\Phi(Ber(1/2)) \approx Q$.

   Lemma 4.2.4. *If $\lambda \le O(\frac{1}{\sqrt{\log n}})$, then $d_{TV}(\Phi(Ber(1/2)), Q) = o(n^{-3})$.*

   Proof sketch for lemma 4.2.4. First note that the support is $(-\infty, \frac{\lambda}{2} + \frac{1}{\lambda}\log 2)$, then direct calculation yields $Z = 1 + o(n^{-3})$. Therefore $d_{TV} = \frac{1}{2}\left(\int |\frac{v(x)+P(x)}{2} - Q(x)|\mathbf{1}_{2Q-P>0}dx + \int |\frac{P(x)}{2} - Q(x)|\mathbf{1}_{2Q-P<0}dx\right) = o(n^{-3})$. □

   Lemma 4.2.5. *If $X_1 \perp\!\!\!\perp X_2$ and $Y_1 \perp\!\!\!\perp Y_2$, then $d_{TV}((X_1,X_2),(Y_1,Y_2)) \le d_{TV}(X_1,Y_1) + d_{TV}(X_2,Y_2)$.*

   Now we construct matrix $W \in \mathbb{R}^{n\times n}$, such that $W_{ij} = W_{ji} = \Phi(E_{ij})$ for $i \ne j$ otherwise 0. By tensorization lemma 4.2.5, the total variation distance between upper triangle of W and X is

   $$d_{TV}(W_{upper}, X_{upper}) \le o(n^2 \cdot n^{-3}) = o(n^{-1}).$$

2. Clone Gaussian variables to symmetrize.

   There is an algorithm to clone independent copies of any gaussian variable.

   Theorem 4.2.3. *Let $X \sim \mathcal{N}(\mu, \sigma^2)$ as input, we can output $\mathcal{N}(\frac{\mu}{\sqrt{2}}, \sigma^2)^{\otimes 2}$.*

   Proof sketch for Theorem 4.2.3. Let $Z \perp\!\!\!\perp X$ with $Z \sim \mathcal{N}(0,\sigma^2)$, then the joint distribution of $\frac{1}{\sqrt{2}}\begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix}\begin{pmatrix}X\\ Z\end{pmatrix}$ is what we want. □

Now we take an antisymmetric matrix A with i.i.d. $\mathcal{N}(0,1)$ random variables below its main diagonal and set $W \leftarrow \frac{1}{\sqrt{2}}(W+Z)$. Then we have the total variation distance between off-diagonal entries of W and X is

$$d_{TV}(W_{off}, X_{off}) = o(n^{-1})$$

.

3. Add diagonal randomness.

   Fill the diagonal entries of W with *i.i.d.* $\mathcal{N}(0,1)$ and permutation the columns randomly. The following techical lemma guarantee the total variation may not be too large.

   Lemma 4.2.6 (lemma 8 in [Brennan et al., 2019])**.** *Given distribution P and Q. Let M be random matrix whose diagnoal entries are i.i.d samples from P and off-diagonal entries are i.i.d samples from Q. Choose a permutation σ on $[n]$ uniformly at random, permutate columns of M to yield M_σ. Then, we have*

$$d_{TV}(M_\sigma, Q^{\otimes n\times n}) \le \sqrt{\frac{\chi^2(P,Q)}{2}}.$$

   Proof sketch for lemma 4.2.6. First, note that $d_{\mathrm{TV}}\left(M_\sigma, Q^{\otimes n\times n}\right) \le \frac{1}{2}\sqrt{\chi^2\left(M_\sigma, Q^{\otimes n\times n}\right)}$.

   Then direct calculation yields

$$\begin{aligned}
&\chi^2\left(M_\sigma, Q^{\otimes n\times n}\right)+1 = \int \frac{\mathbb{E}_\sigma\left[\mathbb{P}_{M_\sigma}(X\mid\sigma)\right]^2}{\mathbb{P}_{Q^{\otimes n\times n}}(X)}dX = \mathbb{E}_{\sigma,\sigma'}\int \frac{\mathbb{P}_{M_\sigma}(X\mid\sigma)\mathbb{P}_{M_{\sigma'}}(X\mid\sigma')}{\mathbb{P}_{Q^{\otimes n\times n}}(X)}dX\\
&= \mathbb{E}_{\sigma,\sigma'}\int\left(\prod_{i:\sigma(i)=\sigma'(i)}\frac{P\left(X_{i\sigma(i)}\right)^2}{Q\left(X_{i\sigma(i)}\right)}\right)\left(\prod_{i:\sigma(i)\neq\sigma'(i)}P\left(X_{i\sigma(i)}\right)\right)\times\left(\prod_{i:\sigma(i)\neq\sigma'(i)}P\left(X_{i\sigma'(i)}\right)\right)\left(\prod_{j\neq\sigma(i),j\neq\sigma'(i)}Q\left(X_{ij}\right)\right)dX\\
&= \mathbb{E}_{\sigma,\sigma'}\prod_{i:\sigma(i)=\sigma'(i)}\left(\int\frac{P\left(X_{i\sigma(i)}\right)^2}{Q\left(X_{i\sigma(i)}\right)}dX_{i\sigma(i)}\right)\\
&= \mathbb{E}_{\sigma,\sigma'}\left(1+\chi^2(P,Q)\right)^{\left|\{i:\sigma(i)=\sigma'(i)\}\right|} \le e^{\chi^2(P,Q)} \le 2\cdot\chi^2(P,Q)+1,
\end{aligned}$$

   where the last inequality is because $e^x \le 1+2x$ for $x\in[0,1]$, and the second to last inequality is because MGF of $Y := |\{i:\sigma(i)=\sigma'(i)\}|$ is upper bounded by MGF of Poisson distribution with rate 1, therefore $\mathbb{E}[e^{tY}] \le e^{e^t-1}$. □

   Now we replace $\mathcal{N}(\lambda,1)$ in diagonal entries of X to $\mathcal{N}(0,1)$ and permutate the columns with the same permutation as W, denote the result matrix by X_{rep}. Then, use similar proof to lemma 4.2.6, assume $\lambda \le O(1/\sqrt{\log n})$ we have

$$d_{TV}(X_{rep}, X) \le O(\sqrt{\chi^2(N(0,1),N(\lambda,1))}) = O(\sqrt{e^{\lambda^2}-1}) = O\left(\frac{1}{\sqrt{\log n}}\right).$$

   Therefore, $d_{TV}(W,X) \le d_{TV}(W,X_{rep})+d_{TV}(X_{rep},X) = d_{TV}(W_{off},X_{off})+d_{TV}(X_{rep},X) = O\left(\frac{1}{\sqrt{\log n}}\right)$.

By reduction steps above, for any k, as long as $\lambda \le O(1/\sqrt{\log n})$ we obtain a random matrix that almost satisfies distribution of $BC(n,k,\lambda)$ from a random graph of $PC(n,k)$. Note that this regime is where $\lambda = n^{-\beta}$ for $\beta \ge 0$.

Therefore, assume the hardness result of $PC(n,k)$ for $k=n^\alpha$ with $\alpha < 1/2$, we obtain the hardness of $BC(n,k,\lambda)$ on the line segment $\{(\alpha,0): 0<\alpha<1/2\}$.

Another computational boundary $\{(\alpha,\beta): 1/2<\alpha<2/3, \beta = 2\alpha-1\}$ can be reduced from $BC(n,n^\alpha,n^{-\beta})$ with $\alpha = 1/2, \beta = 0$. This is done by the following observation: if we clone the entries of X 4 times to obtain a new random matrix $\tilde{X}\in\mathbb{R}^{2n\times 2n}$ with $2k\times 2k$ submatrix, then the resulting problem is actually $BC(2n,2k,\lambda/2)$, since the signal shrinks by a factor of $1/\sqrt{2}$ each time it is cloned.

Surprisingly, note that $\frac{(2k)^2(\lambda/2)}{2n} = \frac{k^2\lambda}{n}$. Therefore, if $BC(n,k,\lambda)$ is at computational boundary, so does $BC(2n,2k,\lambda/2)$. Then, repeat the procedure for l times, we show $BC(2^l n, 2^l k, 2^{-l}\lambda) := BC(\tilde{n},\tilde{k},\tilde{\lambda})$ is at

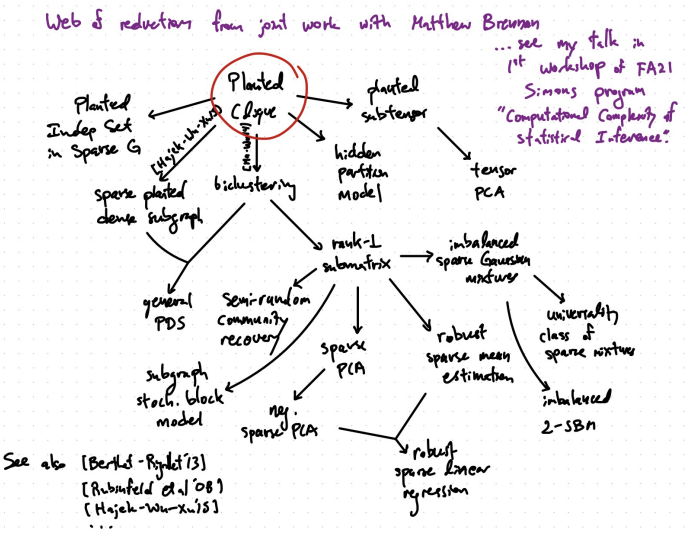


Figure 4.3: Web of reduction xiaobo: cite:from brysler's talk

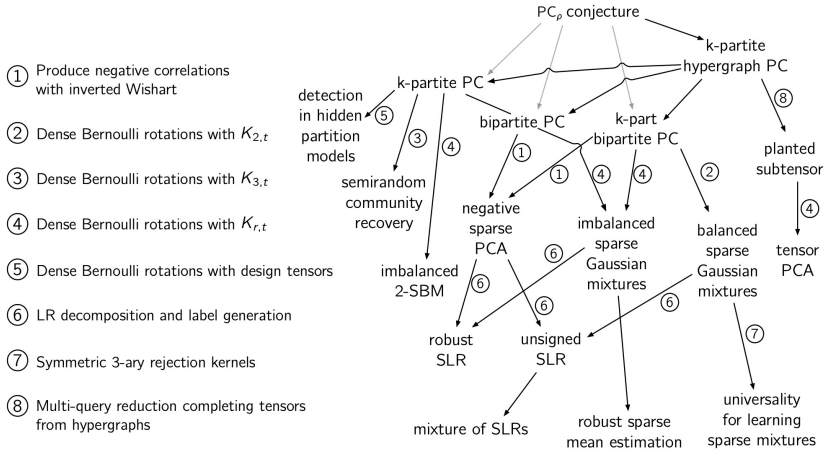


Figure 4.4: Web of reduction [Brennan and Bresler, 2020]

computational boundary. In all, let $\tilde{k} = \tilde{n}^{\tilde{\alpha}}$, $\tilde{\lambda} = \tilde{n}^{-\tilde{\beta}}$, starting from the hardness result at $\alpha = 1/2, \beta = 0$, we actually establish the hardness result in the regime

$$\tilde{\alpha} = \log_{\tilde{n}} \tilde{k} = \frac{l \log 2 + \frac{1}{2} \log n}{l \log 2 + \log n}, \quad \tilde{\beta} = \log_{\tilde{n}} \tilde{\lambda} = \frac{l \log 2}{l \log 2 + \log n}$$

for any $l \geq 0$, which shows the computational boundary at $\tilde{\beta} = 2\tilde{\alpha} - 1$ if we take $l = O(\log n)$.

Further directions

Web of reduction

We actually have a web of reduction to unified many different problems(see figure 4.3). Later in [Brennan and Bresler, 2020] the authors construct a novel problem to reduce more problems with different structures(see figure 4.4). xiaobo: write some intro....

An open problem is to consider the equivalence of the problem, that is, reduction from both sides.

Other related problems and complexities

- Communication complexity

xiaobo: add some intro...rNLA type of papers usually emphrase model of query, but info-comp gap type of papers ususally don't? check that, e.g. what's query model of hardness conjecture of planted cliques

[Rashtchian et al., 2021] attemps to make communication problem(alice and bob) as a **new primitive**, prior to planted clique problem and its variants, under different query models.

For example, a reduction from planted clique(average-case) $PC(n,k)$ to unique disjointness(worst-case) $UD(l)$(which may be too lose since the primitive is nonrandom):

1. Given inputs $x, y \in \{0,1\}^l$ for Alice and Bob respectively. Construct a clique partition of K_n with a collection S of l edge-disjoint k-cliques, which requires $l = \Theta(n^2/k^2)$.
2. Introduce randomness. First randomly picked the edges that aren't covered by S with probability $1/2$, denote the resulting graph by G'. Then, each of the k-cliques K_k^i in S is randomly colored at random with 4 colors, independently of each other.
3. Construct graph. For each $i \in [l]$, if $x_i = 0$, pick all edges in K_k^i with colors 1 or 3 as graph G_1^i, otherwise pick edges with colors 1 or 2. If $y_i = 0$, pick all edges in K_k^i with colors 1 or 4 as graph G_2^i, otherwise pick edges with colors 3 or 4.
4. Let $G = G' \cup (\cup_{i\in[l]} G_1^i \oplus G_2^i)$. Then $G \sim \mathcal{H}_1$ if there is a unique i such that $x_i = y_i = 1$, while $G \sim \mathcal{H}_0$ if there is no such i. Therefore, if we can solve $PC(n,k)$ under some query model(e.g. $\mathbb{F}_2$ sketching), we need at least $\Omega(n^2/k^2)$ queries due to communication complexity.

- Randomized numerical linear algebra

  xiaobo: woodruff's paper usually map problem they care into classical problem(e.g. communication), e.g. Hutch++ [Meyer et al., 2021]. But they don't seem to estabilish system like [Brennan et al., 2019]

- Graph theory

  xiaobo: search cheng mao, jian ding's works, their problems may be reduce other problem, but first we need to abstract their math structures

-

4.2.2 Failure of general algorithms

Sum-of-Squares

This section is mainly based on [Hopkins, 2022a] [Hopkins, 2022b] [Schramm, 2022].

When proving the identifiability of a hypothesis testing problem, it is common to use constructive proof by constructing a statistic that demonstrates the identifiability. However, these statistics may not be computable in polynomial time. For example, the searching statistics $T_{search} = \frac{1}{k}\max_{|S|=|T|=k}\sum_{i\in S, j\in T} X_{ij}$ for biclustering problem 4.2.1 is inefficient[1].

The Sum-of-Squares(SOS) method is a novel framework to prove, with construction of poly-time computable instances, identifiability of statistical problems. By this way, we simultaneously obtain a poly-time algorithm to solve the problem once we prove the identifiability.

More precisely, SOS is a **system** for proving problems that non-negativity of $p_1, ..., p_m$ implies non-negativity of p. Denote it by $p_1(x) \geq 0, ..., p_m(x) \geq 0 \vdash p(x) \geq 0$. We call a SOS proof is d**-degree** if $\max\{\max_{i\in[m]} deg(p_i), deg(p)\} = d$, denoted by $\vdash_d$.

The framework is

1. Translation

   Find some (low) degree-d n-variable polynomials $p_1(x), p_2(x), ..., p_m(x)$ and $p(x)$, $x = (x_1, ..., x_n) \in \mathbb{R}^n$. Translate the assumptions of the problem into x s.t. $p_1(x) \geq 0, ..., p_m(x) \geq 0$[2], and the objects of the

[1] We usually care about time-efficiency, rather than space-efficiency. See [Schramm, 2022] lecture 0.

[2] Actually, we may WLOG relax equality constraint to inequality constraint, since $p = q$ is equivalent to $p - q \geq 0$ and $q - p \geq 0$.

problem into $p(x) \geq 0$. Then the statement to prove for the problem is **translated** to

$$p_1(x) \geq 0, ..., p_m(x) \geq 0 \vdash_d p(x) \geq 0. \tag{4.2}$$

2. Verification

   Starting from some basic SOS inequalities. Use language of SOS proof to **verify** (4.2).

   (a) (Basic inequalities) To prove "$\vdash$", we usually construct a decomposition such that

   $$p(x) = \sum_{S \subseteq [m]} \prod_{i \in S} p_i(x) \cdot q_S(x) + g(x),$$

   where q_S, g are sum-of-squares decompositions $q_S(x) = \sum_j q_{S,j}^2(x)$ and $g(x) = \sum_j g_j^2(x)$. This is to say, we use sum-of-squares polynomials to prove non-negativity arguments. By this way, if $p - q$ is sum-of-squares, we also call the inequalities **"SOS less than"**, and denote them by "$q \preccurlyeq p$" instead.

   Note that, inequality of polynomial $\mathbb{R}[x] \ni p \leq q \in \mathbb{R}[x]$ may not always have SOS of polynomial, *i.e.* $q - p = SOS$, but can be written as SOS of rational function(see [Schramm, 2022] lecture 0).

   (b) (Steps of induction) Then, we operate the inequalities in this proof system.

   For example, $a(x) \geq 0 \vdash b(x) \geq 0$, then $a(x)p^2(x) \geq 0 \vdash b(x)p(x) + q^2(x) \geq 0$.

3. Solution

   Once (4.2) is verified, the statement of the SOS system yields a poly-time algorithm by the main theorem 4.2.4.

   Theorem 4.2.4. *There is a $(n \cdot m)^{O(d)}$ time algorithm to solve SOS system (4.2).*

   The proof is left to further sections.

Example: robust mean estimation

Problem 4.2.3. *Given distribution $\mathcal{D}$ on $\mathbb{R}^d$ and vector $\mu \in \mathbb{R}^d$, such that $\mathbb{E}_{X \sim \mathcal{D}}[X] = \mu$ and $\mathbb{E}_{X \sim \mathcal{D}}(X - \mu)(X - \mu)^T = \Sigma \preccurlyeq C \cdot I$ for some constant $C > 0$. An adversary modify ϵ-proportion of samples $X_1', ..., X_n' \overset{\text{i.i.d.}}{\sim} \mathcal{D}$, yielding corrupted samples $X_1, ..., X_n$ as observation. Use corrupted samples to estimate μ with small error.*

Here we want to know the **identifiability** of this problem. That is, whether it is possible to estimate μ from the ϵ-corruption observation or not. This is in essence a signal-noise tradeoff problem, and we first formulate a sufficient condition of identifiability by classical way, that is:

Theorem 4.2.5. *If*

1. *(Strong signal) $n \gg poly(d/\epsilon)$,*
2. *(Weak noise) there exists an $(1 - \epsilon)n$-subset $S \subseteq [n]$ such that, the empirical covariance matrix over $\{X_i\}_{i \in S}$ is bounded,*

then the problem is identifiable. Moreover, there exists estimator $\hat{\mu}$ s.t. $\|\mu - \hat{\mu}\| \leq O(\sqrt{\epsilon})$.

We will see, the classical way of proof can be constructive and yield an algorithm to calculate $\hat{\mu}$ if the problem is identifiable. However, **the running time for outputting $\hat{\mu}$ may be exponentially long**. On the other hand, SOS proof yields poly-time algorithm. Assume $n \gg poly(d/\epsilon)$, we have

- Classical proof: $O(\binom{n}{n\epsilon})$ time.
- SOS proof: $[d(d+n)]^{O(1)}$ time.

Proof for Theorem 4.2.5. 1. First, we show **how strong signal** will the **true** samples give us.

Let $\bar{X}' = \frac{1}{n}\sum_{i=1}^n X_i'$ and $\Sigma_X' = \frac{1}{n}\sum_{i=1}^n (X_i' - \bar{X}')(X_i' - \bar{X}')^T$, the concentration results yields

Lemma 4.2.7. *For any $\epsilon > 0$, if $n = \Omega(poly(d/\epsilon))$, we have*

(a) $\|\mu - \bar{X}'\| \le \sqrt{\epsilon}$ *w.h.p,*

(b) $\|\Sigma'_X - \Sigma\| \le C$ *w.h.p for some constant $C > 0$, and thus $\Sigma'_X \preccurlyeq C \cdot I$ w.h.p.*

Proof sketch for lemma 4.2.7. (a) For any constant $C_1 > 0$,

$$\mathbb{P}(\|\mu - \bar{X}'\| \ge C_1\sqrt{\epsilon}) \le \mathbb{P}\left(\exists j \in [d] \text{ s.t. } \left|\mu_j - \frac{1}{n}\sum_{i=1}^{n} X'_{i,j}\right| \ge C_1\sqrt{\frac{\epsilon}{d}}\right) \le \sum_{j=1}^{d} \frac{O_n(C/n)}{C_1^2\epsilon/d} = O_n\left(\frac{d^2}{\epsilon}\cdot\frac{1}{n}\right).$$

(b) Since $\Sigma'_X = \frac{1}{n}\sum_{i=1}^{n}(X'_i - \mu)(X'_i - \mu)^T - (\mu - \bar{X}')(\mu - \bar{X}')^T$. By proof above we know $\|(\mu - \bar{X}')(\mu - \bar{X}')^T\| = \|\mu - \bar{X}'\|^2 \le \epsilon$ *w.h.p.* Then, under some regularity conditions, we have $\left\|\frac{1}{n}\sum_{i=1}^{n}(X'_i - \mu)(X'_i - \mu)^T - \Sigma\right\| \le \epsilon$ *w.h.p*, and thus $\|\Sigma'_X\| \le \|\Sigma\| + 2\epsilon \le C$ *w.h.p* for some constant $C > 0$.

For example, if we assume boundedness of $\|X'_1 - \mu\|$, by matrix bernstein inequalities we have

$$\mathbb{P}\left(\left\|\frac{1}{n}\sum_{i=1}^{n}(X'_i - \mu)(X'_i - \mu)^T - \Sigma\right\| \ge \epsilon\right) \le 2d\exp\left\{-O\left(\frac{n\epsilon^2}{\|\Sigma\| + \epsilon}\right)\right\}.$$

Note that there exists other condition such as boundedness of the fourth moment.

□

This means, when n is sufficiently large, **signal** from empirical mean of true samples is **strong** enough for estimation. However, the true samples are contaminated and some information is lost.

2. Then, we control the **noise** due to ϵ-corruption.

Note that, ϵ-corruption means ϵ-total variation distance of empirical distributions of true and corrupted samples. That is, $d_{TV}(\mathcal{D}'_X, \mathcal{D}_X) \le \epsilon$. Intuitively, if the variance are bounded, then small total variation distance implies closeness of the means, since **boundedness of variance implies the differences between two distribution are not so wild** and thus **the noise is weak**. Therefore, empirical mean $\bar{X}$ of corrupted samples X may be close to empirical mean $\bar{X}'$ of true samples X'.

Formally,

Lemma 4.2.8. *Let X, Y be random vectors on $\mathbb{R}^d$, such that $\mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^T \preccurlyeq C \cdot I$ and $\mathbb{E}(Y - \mathbb{E}Y)(Y - \mathbb{E}Y)^T \preccurlyeq C \cdot I$ for some constant $C > 0$. If $d_{TV}(X, Y) \le \epsilon$, then $\|\mathbb{E}X - \mathbb{E}Y\| \le O(\sqrt{\epsilon})$.*

Proof sketch for lemma 4.2.8. By lemma 4.2.2, take coupling over (X, Y) such that $\mathbb{P}(X \ne Y) \le \epsilon$. Then,

$$\begin{aligned}\|\mathbb{E}X - \mathbb{E}Y\|^2 &= \langle \mathbb{E}(X - Y)\mathbf{1}_{X\ne Y}, \mathbb{E}X - \mathbb{E}Y\rangle = \mathbb{E}\mathbf{1}_{X\ne Y}\langle X - Y, \mathbb{E}X - \mathbb{E}Y\rangle \\ &\overset{(a)}{\le} \sqrt{\mathbb{E}\mathbf{1}^2_{X\ne Y}}\sqrt{\mathbb{E}\langle X - Y, \mathbb{E}X - \mathbb{E}Y\rangle^2} \le \sqrt{\epsilon}\cdot\sqrt{\mathbb{E}\langle X - Y, \mathbb{E}X - \mathbb{E}Y\rangle^2},\end{aligned} \tag{4.3}$$

where (a) uses Cauchy-Schwarz inequality. Then,

$$\begin{aligned}&\mathbb{E}\langle X - Y, \mathbb{E}X - \mathbb{E}Y\rangle^2 = \mathbb{E}\left[\langle X - \mathbb{E}X, \mathbb{E}X - \mathbb{E}Y\rangle - \langle Y - \mathbb{E}Y, \mathbb{E}X - \mathbb{E}Y\rangle + \langle \mathbb{E}X - \mathbb{E}Y, \mathbb{E}X - \mathbb{E}Y\rangle\right]^2 \\ &\overset{(b)}{\le} O(1)\cdot\left[\mathbb{E}\langle X - \mathbb{E}X, \mathbb{E}X - \mathbb{E}Y\rangle^2 + \langle Y - \mathbb{E}Y, \mathbb{E}X - \mathbb{E}Y\rangle^2 + \langle \mathbb{E}X - \mathbb{E}Y, \mathbb{E}X - \mathbb{E}Y\rangle^2\right] \\ &\overset{(a)}{O}(1)\cdot\left[\mathbb{E}\|X - \mathbb{E}X\|^2\cdot\|\mathbb{E}X - \mathbb{E}Y\|^2 + \mathbb{E}\|Y - \mathbb{E}Y\|^2\cdot\|\mathbb{E}X - \mathbb{E}Y\|^2 + \|\mathbb{E}X - \mathbb{E}Y\|^4\right] \\ &\overset{(c)}{\le} O(1)\cdot[\|\mathbb{E}X - \mathbb{E}Y\|^2 + \|\mathbb{E}X - \mathbb{E}Y\|^4],\end{aligned} \tag{4.4}$$

where (b) uses $2ab \le a^2 + b^2$ and (c) uses boundedness of variance. Therefore, we have

$$\begin{aligned}&\|\mathbb{E}X - \mathbb{E}Y\| \le O(\sqrt{\epsilon})\sqrt{1 + \|\mathbb{E}X - \mathbb{E}Y\|^2} \le O(\sqrt{\epsilon})(1 + \|\mathbb{E}X - \mathbb{E}Y\|) \\ \Longrightarrow &\|\mathbb{E}X - \mathbb{E}Y\| \le O(\sqrt{\epsilon}).\end{aligned} \tag{4.5}$$

□

3. Finally, it suffices to show empirical covariance Σ_X of corrupted samples is bounded, since we already show empirical covariance Σ'_X of true samples is bounded in lemma 4.2.7, and $d_{TV}(\mathcal{D}'_X, \mathcal{D}_X)$ is small. However, since the modification can be very wild on the corrupted sample, we may not have $\Sigma_X \preccurlyeq O(1) \cdot I$.

   Fortunately, the **assumption** that $\Sigma_S = \frac{1}{|S|}\sum_{i\in S}(X_i - X_S)(X_i - X_S)^T \preccurlyeq C \cdot I$ for some constant $C > 0$, where $X_S = \frac{1}{|S|}\sum_{1\in S} X_i$, reduces the wildness of the corruption. Then we can use X_S as an estimator instead of $\bar{X}'$. It suffices to show empirical distirbution $\mathcal{D}_S$ over $\{X_i\}_{i\in S}$ are not so far from $\mathcal{D}'_X$ *w.r.t.* total variation distance.

   Indeed, we have $d_{TV}(\mathcal{D}'_X, \mathcal{D}_S) \le d_{TV}(\mathcal{D}'_X, \mathcal{D}_X) + d_{TV}(\mathcal{D}_X, \mathcal{D}_S) \le 2\epsilon$. Then we apply lemma 4.2.8 to distribution $\mathcal{D}'_X$ and $\mathcal{D}_S$ to yield $\|\bar{X}' - X_S\| \le O(\sqrt{\epsilon})$ *w.h.p.*, and thus $\|X_S - \mu\| \le O(\sqrt{\epsilon})$ *w.h.p.*.

□

This constructive proof yields a naive algorithm to calculate $\hat{\mu}$.

Algorithm 3 Naive algorithm for robust mean estimation

Input: ϵ-corruption Samples $X_1, ..., X_n$.
Output: Mean estimation $\hat{\mu}$ such that $\|\hat{\mu} - \mu\| \le O(\sqrt{\epsilon})$.

```
for S ⊆ [n] with |S| = (1 − ε)n do
    Calculate X_S = (1/|S|) Σ_{1∈S} X_i and Σ_S = (1/|S|) Σ_{i∈S} (X_i − X_S)(X_i − X_S)^T.
    if ‖Σ_S‖ ≤ 2 then
        Return μ̂ = X_S
    end if
end for
```

However, the algorithm is inefficient, since the **running time of brutal search is** $O(\binom{n}{n\epsilon})$. Surprisingly, with some translation from this direct proof, we can reformulate the problem into SOS framework. The constructive proof of SOS yields a **poly-time** computable algorithm to calculate $\hat{\mu}$.

Proof for SOS version of Theorem 4.2.5.

1. Translation

   First we translate the assumptions into non-negativity of multivariate polynomials.

   Introduce variables $y \in \mathbb{R}$, $y_1, ..., y_n \in \mathbb{R}^d$, indicator variables $w_1, ..., w_n$ and auxiliary variables $B \in \mathbb{R}^{d\times d}$. From the previous analysis, we need these variables satisfying constraint below, denoted by $\mathcal{A}$:

   (a) (ϵ-corruption) Find an $(1-\epsilon)n$-subset of corrupted samples.
   - $w_i^2 = w_i$ for $i \in [n]$,
   - $w_i X_i = w_i y_i$ for $i \in [n]$,
   - $\sum_{i=1}^n w_i = (1-\epsilon)n$.

   (b) (Bounded variance) Corruption on that subset is not so wild.
   - $C \cdot I - Cov(y_1, ..., y_n) = BB^T$ for some constant $C > 0$.

   (c) (Estimator) Another variable for estimation
   - $y = \frac{1}{n}\sum_{i=1}^n y_i$.[3]

   There are $n + d\cdot n + 1 + d\cdot d + 1$ equations of polynomials in total. For each equation $p = q$, we actually obtain two inequalities $p - q \ge 0$ and $q - p \ge 0$. Therefore the original assumptions now become $p_1(x) \ge 0, ..., p_m(x) \ge 0$ with $x = (y, y_1, ..., y_n, w_1, ..., w_n, B) \in \mathbb{R}^{1+(d+1)n+d^2}$ and $m = 2[(d+1)n+d^2+2]$.

   Similarly, the objects of the problem become $p(x) = O(\epsilon) - \|y - \mu\|_2^2 \ge 0$. Therefore the statement is

$$p_1(x) \ge 0, ..., p_m(x) \ge 0 \vdash p(x) \ge 0 \quad \text{or} \quad \mathcal{A} \vdash p(x) \ge 0. \tag{4.6}$$

[3] Though it's slightly different from original condition(only consider samples over subset S), letting $y_i = 0$ for i s.t. $w_i = 0$ we obtain the same modeling as classicl proof there. Actually, both settings are equivalent.

2. Verification

   Then, prove (4.6).

   (a) (Basic inequalities) From the proof of lemma 4.2.8, we only need the following two inequalities

   - $(a+b+c)^2 \preccurlyeq O(1)\cdot(a^2+b^2+c^2)$.

     This is because $a^2+b^2-2ab=(a-b)^2$, then we have $2ab \preccurlyeq a^2+b^2$(or $2ab \preccurlyeq a^2+b^2$), and then $(a+b)^2 \preccurlyeq 2(a^2+b^2)$(or $(a+b)^2 \preccurlyeq 2(a^2+b^2)$).

   - (Cauchy-Schwarz) $\left(\sum_{i=1}^n a_i b_i\right)^2 \preccurlyeq \left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right)$.

     This is because $\left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right)-\left(\sum_{i=1}^n a_i b_i\right)^2=\sum_{i<j}^n (a_i b_j - a_j b_i)^2$.

   (b) (Steps of induction) Let X satisfies empirical distribution of X_i, denoted by $\mathcal{D}_X$, X' satisfies empirical distribution of X_i', denoted by $\mathcal{D}'_X$, and Y satisfies empirical distribution of of y_i, denoted by $\mathcal{D}_Y$. Also denote empirical distribution of true smaples X_i' by $\mathcal{D}'_X$. Then all the expectations in the proof of lemma 4.2.8 are polynomials. For example, $\mathbb{E}Y=\frac{1}{n}\sum_{i=1}^n y_i = y$ is degree-1 polynomial.

   If we obtain a feasible solution $x=(y,y_1,...,y_n,w_1,...,w_n,B)$ satisfying all the constraint in the step 1, then we already have $Cov(y_1,...,y_2)\preccurlyeq 2I$ and $d_{TV}(\mathcal{D}_X,\mathcal{D}_Y)\le\epsilon$. As long as $n\gg poly(d/\epsilon)$ for concentration of X_i, all the conditions of lemma 4.2.8 are satisfied, and we further have $d_{TV}(\mathcal{D}'_X,\mathcal{D}_Y)\le 2\epsilon$ and $Cov(X_1',...,X_n')\preccurlyeq C\cdot I$. Now it suffices to modify its proof into SOS framework.

   By (4.3) we have

   $$\|\mathbb{E}X'-\mathbb{E}Y\|^4 \preccurlyeq \epsilon\cdot\mathbb{E}\langle X'-Y,\mathbb{E}X'-\mathbb{E}Y\rangle^2 \quad w.h.p.(\text{ of } X_i'). \tag{4.7}$$

   By (4.4) we have

   $$\mathbb{E}\langle X'-Y,\mathbb{E}X'-\mathbb{E}Y\rangle^2 \preccurlyeq O(1)\cdot[\|\mathbb{E}X'-\mathbb{E}Y\|^2+\|\mathbb{E}X'-\mathbb{E}Y\|^4] \quad w.h.p.(\text{ of } X_i'). \tag{4.8}$$

   Therefore, $\|\mathbb{E}X'-\mathbb{E}Y\|^4 \preccurlyeq O(\epsilon)\cdot\|\mathbb{E}X'-\mathbb{E}Y\|^2$ $w.h.p.$, and thus we have $\|\bar{X}'-y\|^4\preccurlyeq O(\epsilon)\|\bar{X}'-y\|^2$ $w.h.p.$, where $\bar{X}'=\frac{1}{n}\sum_{i=1}^n X_i'$.

   Note that, max degree of all the polynomials appeared in the proof above is 6, thus it is a degree-6 SOS proof. Therefore we have

   $$\mathcal{A}\vdash_6 \|\bar{X}'-y\|^4\preccurlyeq O(\epsilon)\|\bar{X}'-y\|^2 \quad w.h.p.(\text{ of } X_i'). \tag{4.9}$$

   Though the RHS is slightly different from $p(x)$ defined in (4.6), the two polynomial are equivalent, since $\|\mu-\bar{X}'\|\le\sqrt{\epsilon}$ $w.h.p.$ by concentration.

3. Solution

   Finally, note that $m=2[(d+1)n+d^2+2]=O(d(d+n))$ and dimension of x is $1+(d+1)n+d^2=O(d(d+n))$. By theorem 4.2.4, as long as $n=poly(d/\epsilon)$, we only need **at most** $[d(d+n)]^{O(1)}$ **running time** to solve (4.9). Later we will see the detail implementation by SDP.

□

Rigorous proof of main theorem 4.2.4

Detailed implementation of SOS algorithm is based on a definition called pseudoexpectation

Definition 4.2.3 (pseudoexpectation [Schramm, 2022]). *For a set of polynomial axioms $\mathcal{A}=\{p_1\ge 0,...,p_m\ge 0\}$, we say that $\widetilde{\mathbb{E}}:\mathbb{R}[x]\to\mathbb{R}$ is a degree-d pseudoexpectation satisfying $\mathcal{A}$ if it is a linear operator with the following properties:*

1. *Scaling:* $\widetilde{\mathbb{E}}[1]=1$
2. *Respecting axioms:* $\widetilde{\mathbb{E}}\left[a^2 p_i\right]\geqslant 0$ *for all* $i\in[m]$ *and* $b\in\mathbb{R}[x]$ *satisfying* $\deg\left(a^2p_i\right)\leqslant d$.

3. Non-negativity of squares: $\widetilde{\mathbb{E}}\left[h^2\right] \geqslant 0$ *for any polynomial* $h \in \mathbb{R}[x]$ *with* $\deg(h) \leqslant d/2$

Given a constraint set $\mathcal{A}$, we need to find a pseudoexpectation $\tilde{\mathbb{E}}$ *w.r.t.* $\mathcal{A}$, since it transform the SOS language of inequalities into actual quantities. For example, for any $s \leq d$, if we have $\mathcal{A} \vdash_s p \preccurlyeq q$, then $\tilde{\mathbb{E}}[p] \leq \tilde{\mathbb{E}}[q]$.

In robust mean estimation, by the non-negativity of $\widetilde{\mathbb{E}}$ applied to squares, we have

$$0 \leq \widetilde{\mathbb{E}}\left[\left(\|\bar{X}' - y\|^2 - \widetilde{\mathbb{E}}\left[\|\bar{X}' - y\|^2\right]\right)^2\right] = \widetilde{\mathbb{E}}\left[\|\bar{X}' - y\|^4\right] - \widetilde{\mathbb{E}}\left[\|\bar{X}' - y\|^2\right]^2 \leqslant \widetilde{\mathbb{E}}\left[\|\bar{X}' - y\|^2\right]\left(O(\varepsilon) - \widetilde{\mathbb{E}}\left[\|\bar{X}' - y\|^2\right]\right),$$

which implies that $\widetilde{\mathbb{E}}\left[\|\bar{X}' - y\|^2\right] \leqslant O(\varepsilon)$. Similarly, $\|\bar{X}' - \widetilde{\mathbb{E}}[y]\|^2 \leqslant O(\varepsilon)$. So the quantity $\widetilde{\mathbb{E}}[y]$ computed by our algorithm is a good estimate for $\bar{X}'$ and thus for μ.

To find a pseudoexpectatio *w.r.t.* $\mathcal{A}$, by linearity it suffices to prescribed pseudoexpectation $\widetilde{\mathbb{E}}[x_S]$ for any degree-d monomials $x_S = \prod_{i \in S \subseteq [n], |S| \leq d} x_i$. Therefore, consider a large matrix $Z \in \mathbb{R}^{n^{O(d)} \times n^{O(d)}}$, whose rows and columns are indexed by d-subsets of $[n]$. Then, we may take $\widetilde{\mathbb{E}}[x_S] = Z_{A,B}$, where $A \cup B = S$ is an arbitrary partition of S. To make it well-defined, assume $Z_{\emptyset,\emptyset} = 1$ and $Z_{A,B} = Z_{S,T}$ if $A \cup B = S \cup T$.

Now, we need to make sure non-negativity and axiom respecting, The way to find such matrix Z is SDP, problem 4.2.4.

Problem 4.2.4 ([Schramm, 2022]). *Given matrix* $C, P_i \in \mathbb{R}^{r \times r}$, $i \in [m]$, *solve the following optimization problem*

$$\begin{aligned} \max_{Z} \quad & \langle Z, C \rangle \\ s.t. \quad & Z \succeq 0 \\ & \langle Z, P_i \rangle \geq 0 \quad \forall i \in [m] \end{aligned} \tag{4.10}$$

Set $C = 0$, *the problem become finding a feasible solution of the constraints.*

The state-of-art SDP algorithm, such as interior point method and epllisoid method, will solve this system in time $poly(r, m)$.

The steps to find a suitable Z as basis of pseudoexpectation by SDP are as follows

1. First, transform the scalar constraint in $\mathcal{A}$ into matrix inner-produt.

   Indeed, for any $i \in [m]$, $\widetilde{\mathbb{E}}[p_i] \geq 0$ is actually $\langle Z, P_i \rangle \geq 0$, where $(P_i)_{S,T} = \hat{p_i}(S \cup T)$, which is the coefficient of monomials $x_{S \cup T}$ in p_i. To make $\widetilde{\mathbb{E}}[a^2 p_i] \geq 0$ for all $a(x) \in \mathbb{R}_{\leq d/2}[x]$, it suffices to consider $\widetilde{\mathbb{E}}[x_S^2 p_i] \geq 0$ for all $S \subset [n], |S| = d/2$. Therefore, we transform m equality constraint in $\mathcal{A}$ into $m \cdot n^{O(d)}$ equality constraint in SDP.

2. Then, the non-negativity is trivial since $0 \preccurlyeq Z$ and $\widetilde{\mathbb{E}}[h^2] = \hat{h}^T Z \hat{h} \geq 0$, where $\hat{h}$ is vector of coefficient of all monomials in h.

3. Finally, use state-of-art SDP algorithm, we will solve the SOS algrotihm in time $poly(n^{O(d)}, m \cdot n^{O(d)}) = (n \cdot m)^{O(d)}$.

Limitation

1. Are polynomials enough?

   Not all the problems can be translated into non-negativity of polynomials. (LDP has similar issues.) For example, calculate $\frac{Av}{\|Av\|}$ in power method.

2. Inequalities need SOS version

   Not all polynomial inequalities $p \geq q$ has SOS formulation $p - q = \sum_i g_i^2$. But this is significant for efficiency of SOS, since we need to construct pseudoexpectation inequality $\tilde{\mathbb{E}}[p] \geq \tilde{\mathbb{E}}[q]$ by SDP. Time complexity of the SDP exponentially depends on dimension of function space of g_i, since we constrain $\tilde{\mathbb{E}}[h^2] \geq 0 \ \forall h \in \mathbb{R}[x]$. Space of low degree polynomials is low-dimensional.

Hardness and lower bound

xiaobo: see [Barak et al., 2016] and more

More problems

xiaobo: see xiechuhan's notes and more sos papers and [Barak et al., 2016]

Relatioship with low degree polynomial method

xiaobo: see [Hop18],[KWB19] and more

4.2.3 Phenomena in hard regime

Chapter 5

Math Foundation

Basic math tools.

5.1 Key elements

5.1.1 Yao's principle

When we analyze complexity of average-case problem with randomized algorithm, we can analyze a problem with randomized input and deterministic algorithm instead.

Theorem 5.1.1 (Yao's minimax principle [Yao, 1977]). *Given a problem $\mathcal{P}$, let $\mathcal{A}$ be the set of all deterministic algorithms to solve $\mathcal{P}$ and $\mathcal{X}$ be the set of all instances of $\mathcal{P}$. Let $c(a,x)$ be the cost of algorithm $a \in \mathcal{A}$ solving instance $x \in \mathcal{X}$. For any distribution p over $\mathcal{A}$ and any distribution q over input space $\mathcal{X}$, consider randomized algorithm $A \sim p$ and random instance $X \sim q$. Then, we have*

$$\max_{x \in \mathcal{X}} \mathbf{E}[c(A,x)] \geq \min_{a \in \mathcal{A}} \mathbf{E}[c(a,X)]$$

That is, the worst-case expected cost of the randomized algorithm is at least the expected cost of the best deterministic algorithm against input distribution q.

Proof for Theorem 5.1.1. Let $C = \max_{x \in \mathcal{X}} \mathbf{E}[c(A,x)]$ and $D = \min_{a \in \mathcal{A}} \mathbf{E}[c(a,X)]$. We have

$$C = \sum_x q_x C \geq \sum_x q_x \mathbf{E}[c(A,x)] = \sum_x q_x \sum_a p_a c(a,x) = \sum_a p_a \sum_x q_x c(a,x) = \sum_a p_a \mathbf{E}[c(a,X)] \geq \sum_a p_a D = D$$

□

5.1.2 Model of computation

5.2 Concentration and anti-concentration inequalities

5.2.1 Overlap of a vector on a large set

If two subsets of $\mathbb{R}^n$ are "large", then the "overlap" of them is large. Here, "large" means the subsets have large meassure under Gaussian measure, and "overlap" means: For any non-empty $A, B \subseteq \mathbb{R}^n$. Denote by $\gamma_{|A \times B}$ the probability measure corresponding to the normalized restriction of $\mathcal{N}(0, I_n) \times \mathcal{N}(0, I_n)$ to $A \times B$ and let

$$v(A,B) := \mathbb{E}_{(x,y) \sim \gamma_{|A \times B}}[|\langle x, y \rangle|^2].$$

Formally, we have

Theorem 5.2.1 ([Vidick, 2012]). *For any $\eta > 0$, there exists [1] a $\delta > 0$ such that for all large enough n, if A, B both have measure $\gamma(A), \gamma(B) \geq e^{-\delta n}$ then*

$$v(A,B) \geq (1-\eta) v\left(\mathbb{R}^n, \mathbb{R}^n\right) = (1-\eta) n$$

It's continuous version of (4.1), which is more general to yield corruption bound for communication problem with continuous domain $\mathcal{X}, \mathcal{Y}$ and rectangle R.

There is an similar anti-concentration result in [Chakrabarti et al., 2012], where they study ORT problem.

Theorem 5.2.2 (Anti-concentration lemma 4.1 in [Chakrabarti et al., 2012]). *For sufficiently large n and $b \geq 66$, let $\epsilon = 1 - \Phi(2.01b)$. Then there exists $\delta > 0$ such that for all $A, B \subseteq \{\pm 1\}^n$ with $\min\{|A|, |B|\} \geq 2^{n-\delta n}$, we have*

$$\mathbb{P}_{(X,Y)\sim Unif\{A\times B\}}[|\langle X, Y\rangle| > b\sqrt{n}] \geq \epsilon.$$

Remark 5.2.1. *This lemma incorporates lemma 3.3 in [Sherstov, 2012](key of theorem 4.1.2) as a speacial case of $b = 1/4$. Here b is general but the analysis in [Sherstov, 2012] fails when $b > 1$.*

5.2.2 Hutchinson's trace estimator

A simple estimator of trace of matrix by monte carlo method and we use only matrix-vector product oracle.

Definition 5.2.1 (Hutchinson's Estimator). *Given a matrix $A \in \mathbb{R}^{d\times d}$. We estimate $tr(A)$ by*

$$H_m(A) = \frac{1}{m}\sum_{i=1}^{m} g_i^T A g_i = \frac{1}{m} tr(G^T A G),$$

where $G = [g_1, ..., g_m] \in \mathbb{R}^{d\times m}$ is a Gaussian matrix with i.i.d. $\mathcal{N}(0, 1)$ entries.

Theorem 5.2.3 (Hutchinson analysis [Meyer et al., 2021]). *Let $A \in \mathbb{R}^{d\times d}, \delta \in (0, 1/2], \ell \in \mathbb{N}$. Let $\mathrm{H}_\ell(A)$ be the ℓ-query Hutchinson estimator defined above, implemented with mean 0, i.i.d. sub-Gaussian random variables with constant sub-Gaussian parameter. For fixed constants c, C, if $\ell > c\log(1/\delta)$, then with probability $\geq 1 - \delta$,*

$$|\mathrm{H}_\ell(A) - \mathrm{tr}(A)| \leq C\sqrt{\frac{\log(1/\delta)}{\ell}}\|A\|_F.$$

So, if $\ell = O\left(\frac{\log(1/\delta)}{\varepsilon^2}\right)$ then, with probability $\geq 1 - \delta, |\mathrm{H}_\ell(A) - \mathrm{tr}(A)| \leq \varepsilon\|A\|_F$.

Proof sketch for Theorem 5.2.3. First, vectorize gaussian matrix G to $\bar{g} \in \mathbb{R}^{dl}$ and let $\bar{A} = diag\{A, A, ..., A\} \in \mathbb{R}^{dl\times dl}$.

Then, $H_l(A) = \bar{g}^T \bar{A}\bar{g}/l$ and use Hanson-Wright inequality [Vershynin, 2019]. □

5.2.3 Talagrand's inequality

Theorem 5.2.4 (Talagrand [Sherstov, 2012]). *For a fixed convex set $S \subseteq \mathbb{R}^n$ and a random $x \in \{-1, +1\}^n$,*

$$\mathbb{P}[x \in S]\mathbb{P}[\rho(x, S) > t] \leq \mathrm{e}^{-t^2/16},$$

where $\rho(x, S) = \inf_{y\in S}\|x - y\|$.

Intuitively, it means measure concentration around convex hull of large subset $S \subseteq \{\pm 1\}^n$, where the measure means counting measure(or uniform distribution). It can be seen as an isoperimetric inequality in discrete structures.

The usefulness of Talagrand's inequality is led by appropriately choosing set S, according to property we interest. Here are some useful corollaries.

Corollary 5.2.1 ([Sherstov, 2012]). *For every linear subspace $V \subseteq \mathbb{R}^n$ and every $t > 0$, one has*

$$\mathop{\mathbb{P}}_{x\in\{-1,+1\}^n}\left[\left|\|\mathrm{proj}_V x\| - \sqrt{\dim V}\right| > t\right] < 4\mathrm{e}^{-ct^2},$$

where $c > 0$ is an absolute constant.

Proof for Theorem 5.2.1. First, note that it suffices to prove

$$\mathbf{P}[|\rho(x,V)-\sqrt{n-\dim V}|>t]\le 4\mathrm{e}^{-\Omega(t^2)}$$

and $n-\dim V = tr(\mathbb{E}[x^T(I-P_V)x]) = \mathbb{E}[\rho(x,V)^2]$, where P_v is orthogonal projection on V. Therefore, take $S=\{v\in\mathbb{R}^n:\rho(v,V)\le a\}$, then by Talagrand's inequality we have

$$\mathbb{P}[\rho(x,V)\le a]\mathbb{P}[\rho(x,V)>a+t]\le\mathbb{P}[\rho(x,V)\le a]\mathbb{P}[\rho(x,S)>t]\le\mathrm{e}^{-t^2/16}.$$

Then, take $a=median(\rho(x,V)):=m$, we have

$$\mathbb{P}[\rho(x,V)>m+t]\le 2\mathrm{e}^{-t^2/16},$$
$$\mathbb{P}[\rho(x,V)\le m-t]\le 2\mathrm{e}^{-t^2/16}.$$

Finally, the result follows $m=\sqrt{\mathbb{E}[\rho(x,V)^2]}+O(1)$. See [Tao, 2009] for details. □

5.2.4 Sketching

Definition 5.2.2 (projection-cost-preserving [Musco and Musco, 2020])**.** *A matrix $\tilde{A}\in\mathbb{R}^{n\times m}$ is an (ϵ,c,k) projection-cost-preserving sketching of $A\in\mathbb{R}^{n\times d}$ if for any orthogonal projection $P\in\mathbb{R}^{n\times n}$ with rank at most k, we have*

$$(1-\epsilon)\|A-PA\|_F^2\le\|\tilde{A}-P\tilde{A}\|_F^2+c\le(1+\epsilon)\|A-PA\|_F^2.$$

Theorem 5.2.5 (simple sketching is projection-cost-preserving [Musco and Musco, 2020])**.** *Let $S\in\mathbb{R}^{d\times m}$ be a i.i.d. standard Gaussian matrix. If $m\ge c(k+\log(1/\delta))/\epsilon^2$ for some large constant c, we have $\frac{1}{\sqrt{m}}AS$ is a $(\epsilon,0,k)$-projection-cost-preserving sketching of A with probability at least $1-\delta$.*

Proof sketch for Theorem 5.2.5. □

Theorem 5.2.6 (projection-cost-preserving makes projection approximation [Musco and Musco, 2020])**.** *If $\tilde{A}\in\mathbb{R}^{n\times m}$ is an (ϵ,c,k) projection-cost-preserving sketching of $A\in\mathbb{R}^{n\times d}$, then for any subset $\mathcal{T}$ of all orthogonal projection with rank at most k, if we take a $\tilde{P}\in\mathcal{T}$ such that $\|\tilde{A}-\tilde{P}\tilde{A}\|_F^2\le\gamma\min_{P\in\mathcal{T}}\|\tilde{A}-P\tilde{A}\|_F^2$, we have*

$$\|A-\tilde{P}A\|_F^2\le\frac{1+\epsilon}{1-\epsilon}\gamma\min_{P\in\mathcal{T}}\|A-PA\|_F^2+\frac{(1-\gamma)c}{1-\epsilon}.$$

Proof for Theorem 5.2.6. Direct calculation. □

Theorem 5.2.7 ((Upper Bound on Regression Error [Jiang et al., 2021])**.** *Given $\delta\in(0,1/2)$ and matrices A,B with n rows and $rank(A)=k$. Let $S\in\mathbb{R}^{r\times n}$ be a random matrix with i.i.d. Gaussian $\mathcal{N}(0,\frac{1}{r})$. Let $\widetilde{X}=\arg\min_X\|S(AX-B)\|_F$ and $X^*=\arg\min_X\|AX-B\|_F$. Then, if $r=\Omega(k+\log(1/\delta))$, we have with probability at least $1-\delta$*

$$\|A\widetilde{X}-B\|_F\le O(1)\|AX^*-B\|_F.$$

Remark 5.2.2. *[Clarkson and Woodruff, 2009] shows what $O(1)$ is, while the randomness there is rademacher's variable.*

Proof sketch for Theorem 5.2.7. □

5.3 Statistical distance

Theorem 5.3.1 ([Perry et al., 2018] proposition 5.11)**.** *For any $|\beta|<1$, there exists $\delta>0$ such that the following holds. Let $\mathcal{X}=\{\mathcal{X}_n\}$ be a family of prior distribution of spiked vectors x with $1-\delta\le\|x\|\le1+\delta$. Let Q_n be joint distribution of N i.i.d samples from $\mathcal{N}(0,I_n+\beta xx^T)$ and P_n be joint distribution of N i.i.d samples from $\mathcal{N}(0,I_n)$. Then we have*

$$\mathbb{E}_{P_n}\left[\left(\frac{dQ_n}{dP_n}\right)^2\right]=\mathbb{E}_{x,x'\sim\mathcal{X}_n}\left[(1-\beta^2\langle x,x'\rangle^2)^{-N/2}\right].$$

Proof for Theorem 5.3.1. First, expand LHS directly, we have $\mathop{\mathbb{E}}_{P_n}\left[\left(\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\right)^2\right] = \mathop{\mathbb{E}}_{Q_n}\left[\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\right]$

$$\begin{aligned}\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}(y_1,\ldots,y_N) &= \mathop{\mathbb{E}}_{x'\sim\mathcal{X}}\left[\prod_{i=1}^{n}\frac{\exp\left(-\frac{1}{2}y_i^\top\left(I+\beta x'x'^\top\right)^{-1}y_i\right)}{\sqrt{\det\left(I+\beta x'x'^\top\right)}\exp\left(-\frac{1}{2}y_i^\top y_i\right)}\right]\\ &= \mathop{\mathbb{E}}_{x'}\left[\det\left(I+\beta x'x'^\top\right)^{-N/2}\prod_{i=1}^{N}\exp\left(-\frac{1}{2}y_i^\top\left(\left(I+\beta x'x'^\top\right)^{-1}-I\right)y_i\right)\right].\end{aligned}$$

Then, simplify it by Sherman–Morrison formula, we have $\left(I+\beta x'x'^\top\right)^{-1}-I=\frac{-\beta}{1+\beta\|x'\|^2}x'x'^\top$, and then

$$= \mathop{\mathbb{E}}_{x'}\left[\left(1+\beta\|x'\|^2\right)^{-N/2}\prod_{i=1}^{N}\exp\left(\frac{1}{2}\frac{\beta}{1+\beta\|x'\|^2}\langle y_i,x'\rangle^2\right)\right].$$

Finally, passing to the second moment, we compute

$$\begin{aligned}\mathop{\mathbb{E}}_{P_n}\left[\left(\frac{\mathrm{d}Q_n}{\mathrm{d}P_n}\right)^2\right] &= \mathop{\mathbb{E}}_{x,x'}\left[\left(1+\beta\|x'\|^2\right)^{-N/2}\prod_{i=1}^{N}\mathop{\mathbb{E}}_{y_i\sim\mathcal{N}(0,I+\beta xx^\top)}\exp\left(\frac{1}{2}\frac{\beta}{1+\beta\|x'\|^2}\langle y_i,x'\rangle^2\right)\right]\\ &= \mathop{\mathbb{E}}_{x,x'}\left[\left(1+\beta\|x'\|^2\right)^{-N/2}\prod_{i=1}^{N}\left(1-\frac{\beta}{1+\beta\|x'\|^2}\left(\|x'\|^2+\beta\langle x,x'\rangle^2\right)\right)^{-1/2}\right]\\ &= \mathop{\mathbb{E}}_{x,x'}\left[\left(1-\beta^2\langle x,x'\rangle^2\right)^{-N/2}\right]\end{aligned}$$

Note that, the condition about δ is because here the MGF step requires $\frac{\beta}{1+\beta\|x'\|^2}\left(\|x'\|^2+\beta\langle x,x'\rangle^2\right)<1$. □

Theorem 5.3.2 ([Jiang et al., 2021]). *For distribution $P=\mathcal{N}_k(\mu_1,\Sigma_1)$ and $Q=\mathcal{N}_k(\mu_2,\Sigma_2)$, we have*

$$d_{KL}(P,Q)=\frac{1}{2}\left\{(\mu_2-\mu_1)^T\Sigma_2^{-1}(\mu_2-\mu_1)+tr(\Sigma_2^{-1}\Sigma_1)-\log\frac{det(\Sigma_1)}{det(\Sigma_2)-k}\right\}.$$

Proof sketch for Theorem 5.3.2. □

Theorem 5.3.3 (Data processing inequality [Jiang et al., 2021]). *For random variable X,Y and any function f, we have*

$$d_{KL}(f(X),f(Y))\le d_{KL}(X,Y).$$

Proof sketch for Theorem 5.3.3. □

Remark 5.3.1. *There are also generalized data processing inequality for f-divergence. xiaobo: see chenkun's talk about [Simchowitz et al., 2020]*

5.4 Random matrices

5.4.1 Non-asymptotic theory

In this section, we denote the distribution of random matrices $G\in\mathbb{R}^{n\times n}$ with *i.i.d.* $\mathcal{N}(0,1)$ entries by $\mathcal{N}(n)$.

Theorem 5.4.1 (Remaining randomness [Simchowitz et al., 2020] [Jiang et al., 2021]). *Let $G\sim\mathcal{N}(n)$. Let $W=(G+G^T)/2$. For any sequence of vector queries $v_1,\ldots,v_T$, along with oracles $w_i=Wv_i$. Then, conditioned on these observations, there exists a rotation matrix U, independent of w_i, such that*

$$UWU^T=\begin{bmatrix}Y_1 & Y_2^T\\ Y_2 & \widetilde{W}\end{bmatrix},$$

where Y_1,Y_2 are deterministic and $\widetilde{W}=(\widetilde{G}+\widetilde{G}^T)/2$, where $\widetilde{G}\sim\mathcal{N}(n-T)$.

Proof sketch for Theorem 5.4.1. □

Theorem 5.4.2 (concentration of the largest singular value [Jiang et al., 2021])**.** *Let $G \sim \mathcal{N}(n)$. Then, for any $t \geq 0$ we have*

$$\mathbb{P}[s_{max}(G) \leq 2\sqrt{n} + t] \geq 1 - 2e^{-t^2/2}.$$

Proof sketch for Theorem 5.4.2. □

5.4.2 Asymptotic theory

5.5 Matrix analysis

Lemma 5.5.1 ([Meyer et al., 2021])**.** *For any PSD matrix A, we have $\|A - A_k\|_F \leq \frac{tr(A)}{\sqrt{k}}$.*

Proof for lemma 5.5.1. $LHS^2 = \sum_{i=k+1}^{n} \lambda_i^2 \leq \frac{tr(A)}{k} \sum_{i=k+1}^{n} \lambda_i \leq \frac{tr(A)^2}{k}$. □

Lemma 5.5.2. *Given matrices A, B, we have $\arg\min_X \|AX - B\| = \arg\min_X \|AX - B\|_F = A^+B$ and $\arg\min_Y \|YA - B\| = \arg\min_Y \|YA - B\|_F = BA^+$, where M^+ is Moore–Penrose inverse of matrix M.*

Proof for Theorem 5.5.2. xiaobo: see wiki of Moore–Penrose inverse, which shows AA^+, A^+A are both orthogonal projections □

Bibliography

[Bakshi et al., 2022] Bakshi, A., Clarkson, K. L., and Woodruff, D. P. (2022). Low-rank approximation with $1/\epsilon^{1/3}$ matrix-vector products.

[Bandeira et al., 2022] Bandeira, A. S., Alaoui, A. E., Hopkins, S. B., Schramm, T., Wein, A. S., and Zadik, I. (2022). The franz-parisi criterion and computational trade-offs in high dimensional statistics.

[Barak et al., 2016] Barak, B., Hopkins, S. B., Kelner, J., Kothari, P. K., Moitra, A., and Potechin, A. (2016). A nearly tight sum-of-squares lower bound for the planted clique problem.

[Braverman et al., 2021] Braverman, M., Hazan, E., Simchowitz, M., and Woodworth, B. (2021). The gradient complexity of linear regression.

[Brennan and Bresler, 2020] Brennan, M. and Bresler, G. (2020). Reducibility and statistical-computational gaps from secret leakage.

[Brennan et al., 2019] Brennan, M., Bresler, G., and Huleihel, W. (2019). Reducibility and computational lower bounds for problems with planted sparse structure.

[Bresler, 2022a] Bresler, G. (2022a). Guy bresler - mit - average-case reduction techniques i.

[Bresler, 2022b] Bresler, G. (2022b). Guy bresler - mit - average-case reduction techniques ii.

[Chakrabarti et al., 2012] Chakrabarti, A., Kondapally, R., and Wang, Z. (2012). Information complexity versus corruption and applications to orthogonality and gap-hamming.

[Chakrabarti and Regev, 2012] Chakrabarti, A. and Regev, O. (2012). An optimal lower bound on the communication complexity of gap-hamming-distance.

[Clarkson and Woodruff, 2009] Clarkson, K. L. and Woodruff, D. P. (2009). Numerical linear algebra in the streaming model.

[Dharangutte and Musco, 2021] Dharangutte, P. and Musco, C. (2021). Dynamic trace estimation.

[Hopkins, 2022a] Hopkins, S. (2022a). Sam hopkins - mit - sum of squares methods for statistical problems i.

[Hopkins, 2022b] Hopkins, S. (2022b). Sam hopkins - mit - sum of squares methods for statistical problems ii.

[Jiang et al., 2021] Jiang, S., Pham, H., Woodruff, D. P., and Zhang, Q. R. (2021). Optimal sketching for trace estimation.

[Kushilevitz and Nisan, 1997] Kushilevitz, E. and Nisan, N. (1997). *Communication Complexity*. Cambridge University Press, Cambridge.

[Ma and Wu, 2015] Ma, Z. and Wu, Y. (2015). Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3).

[Meyer et al., 2021] Meyer, R. A., Musco, C., Musco, C., and Woodruff, D. P. (2021). Hutch++: Optimal stochastic trace estimation.

[Musco and Musco, 2020] Musco, C. and Musco, C. (2020). Projection-cost-preserving sketches: Proof strategies and constructions.

[Perry et al., 2018] Perry, A., Wein, A., Bandeira, A., and Moitra, A. (2018). Optimality and sub-optimality of PCA I: Spiked random matrix models. *Annals of Statistics*, 46:2416–2451.

[Rashtchian et al., 2021] Rashtchian, C., Woodruff, D. P., Ye, P., and Zhu, H. (2021). Average-case communication complexity of statistical problems.

[Schramm, 2020] Schramm, T. (2020). Pedagogical talk: Frameworks for information-computation tradeoffs.

[Schramm, 2022] Schramm, T. (2022). Lecture notes - the sum-of-squares algorithmic paradigm in statistics.

[Sherstov, 2012] Sherstov, A. A. (2012). The communication complexity of gap hamming distance.

[Simchowitz et al., 2020] Simchowitz, M., Alaoui, A. E., and Recht, B. (2020). Tight query complexity lower bounds for pca via finite sample deformed wigner law.

[Tao, 2009] Tao, T. (2009). Talagrand's concentration inequality. Weblog entry.

[Vershynin, 2019] Vershynin, R. (2019). *High-Dimensional Probability: An Introduction with Applications in Data Science*.

[Vidick, 2012] Vidick, T. (2012). A concentration inequality for the overlap of a vector on a large set with application to the communication complexity of the gap-hamming-distance problem.

[Woodruff et al., 2022] Woodruff, D. P., Zhang, F., and Zhang, Q. (2022). Optimal query complexities for dynamic trace estimation.

[Yang et al., 2018] Yang, W., Zhao, J., Hulett, J., and Chen, A. (2018). Beyond worst case analysis lecture 2: Largest clique in a random graph.

[Yao, 1977] Yao, A. C.-C. (1977). Probabilistic computations: Toward a unified measure of complexity.

[Yao, 1983] Yao, A. C.-C. (1983). Lower bounds by probabilistic arguments.