

Informational-computational trade-off in randomized numerical linear algebra

Xiaobo Yang

April 2023

Contents

1	Practical Applications	7
1.1	Datasets	7
1.2	Statistical applications	7
1.2.1	Gaussian process	7
2	Theory for application	9
2.1	Matrix-vector multiplication queries	9
2.1.1	Low rank approximation	9
2.1.2	Solve linear system	9
3	Theory for areas	11
3.1	Matrix-vector multiplication queries	11
3.1.1	Trace estimation	11
3.1.2	Eigenvalue estimation	17
4	Theoretic Frameworks	19
4.1	Meta-complexity	19
4.1.1	Communication complexity	19
4.1.2	Spiked wishart matrix testing	22
5	Math Foundation	25
5.1	Key elements	25
5.1.1	Yao's principle	25
5.1.2	Model of computation	25
5.2	Concentration and anti-concentration inequalities	25
5.2.1	Overlap of a vector on a large set	25
5.2.2	Hutchinson's trace estimator	26
5.2.3	Talagrand's inequality	26
5.2.4	Sketching	27
5.3	Statistical distance	27
5.4	Random matrices	28
5.4.1	Non-asymptotic theory	28
5.4.2	Asymptotic theory	29
5.5	Matrix analysis	29

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, \dots, r_m and output product Ar_1, \dots, Ar_m to estimate $\text{tr}(A)$.*

There are a two types of query vectors.

1. (Adaptive) Vectors r_1, \dots, 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, \dots, 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, \dots, r_m are adaptive and with integer entries in $\{-2^b, \dots, 2^b\}$, 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 $2/3$, $(1 - \epsilon)\text{tr}(A) \leq t \leq (1 + \epsilon)\text{tr}(A)$.*

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

$$\begin{aligned} \widehat{\text{tr}(A)} &\geq 2(n + \sqrt{n})(1 - 1/\sqrt{n}) = 2n - 2 \quad \text{if } \langle s, t \rangle \geq \sqrt{n}, \\ \widehat{\text{tr}(A)} &\leq 2(n - \sqrt{n})(1 + 1/\sqrt{n}) = 2n - 2 \quad \text{if } \langle s, t \rangle \leq -\sqrt{n}, \end{aligned}$$

which means we can solve GHD problem by simply compare $\widehat{\text{tr}(A)}$ 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 $m = \Omega\left(\frac{1}{\epsilon(b + \log(1/\epsilon))}\right)$. \square

Theorem 3.1.2 (Lower bound of non-adaptive queries [Meyer et al., 2021]). *Let A be a PSD matrix. If r_1, \dots, 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 $3/4$, $(1 - \epsilon)\text{tr}(A) \leq t \leq (1 + \epsilon)\text{tr}(A)$.*

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 $A = G^T D G \in \mathbb{R}^{d \times d}$, 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 $11/12$,

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

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

$$\begin{aligned} t &\geq (1 - \epsilon/4) \text{tr}(A) \geq (1 - \epsilon/4)^2 d/\epsilon \geq (1 - \epsilon/2) d/\epsilon \quad \text{if } D = D_1, \\ t &\leq (1 + \epsilon/4) \text{tr}(A) \leq (1 + \epsilon/4)^2 (1 - \epsilon) d/\epsilon \leq (1 - \epsilon/2) d/\epsilon \quad \text{if } D = D_2, \end{aligned}$$

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 $N \perp L$ and $L \in \mathbb{R}^{n \times (d-m)}$ with i.i.d $\mathcal{N}(0, 1)$. Then we have $M \stackrel{\text{dist}}{=} G^T D G E_m$, where $E_m = [e_1, \dots, e_m]$ 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 $m = \Omega(\frac{1}{\epsilon})$. □

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, \dots, 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 $1 - \delta$, $(1 - \epsilon) \text{tr}(A) \leq t \leq (1 + \epsilon) \text{tr}(A)$.

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 $\text{tr}(W) \leq 2n$ *w.h.p.* for large n , since $\text{tr}(G) \sim \mathcal{N}(0, n)$. Quantitatively, for any $\delta \in (0, 1)$, take $n = \Omega(\log(1/\delta))$ then $\text{tr}(W) \leq 2n$ 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 $|\text{tr}(W)| \leq 2\sqrt{2}\log(1/\delta)$ since $\text{tr}(W) \sim \mathcal{N}(0, 4\log(1/\delta))$. 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 $t \geq (1-\epsilon)\text{tr}(B) = (1-\epsilon)((C+6)\log^{3/2}(1/\delta) - 2\sqrt{2}\log(1/\delta)) > 6(1+\epsilon)\log^{3/2}(1/\delta)$ with probability at least $1 - 2\delta$,
 - if $B \sim Q$, then $t \leq (1+\epsilon)\text{tr}(B) \leq (1+\epsilon)(6\log^{3/2}(1/\delta) + 2\sqrt{2}\log(1/\delta)) < 6(1+\epsilon)\log^{3/2}(1/\delta)$ 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. \square

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, \dots, 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 $1 - \delta$, $|t - \text{tr}(A)| \leq \epsilon \|A\|_p$.

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 algorithm \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 $\langle a, b \rangle = \text{tr}(AB)$. Now the task is to compare $|\text{tr}(AB)|$ with bn .

In each round, we give a query q_i and compute $q_i A$ and send it to Bob, then Bob compute $q_i AB$ and send it back to Alice. Assume that after $r(n)$ rounds, we got a good trace estimation t through $q_1 AB, \dots, q_{r(n)} AB$. Formally, we have

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

2. Then, assume $a, b \sim \text{Unif}\{\{\pm 1\}^{n^2}\}$ independently, by Holder's inequality for $p \in [1, 2]$ and Markovian's inequality we have $\mathbb{P}_{A,B}(\epsilon \|AB\|_p \geq \epsilon \sqrt{tn^{1/p+1}}) \leq \mathbb{P}_{A,B}(\epsilon \|AB\|_F \geq \epsilon \sqrt{tn^{3/2}}) \leq 1/t, \forall t > 0$. Therefore, take δ', t such that $1/\delta' = t = (\frac{1}{\epsilon^{1/p}} \frac{b}{10})^2$, 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 algorithm \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 $\text{tr}(A)$. 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 $q_1, q_2, \dots, q_{r(n)}$, 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 achieve 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 remaining 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

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

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

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

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, \dots, r_m are adaptive, then we need at least,*

- (bit complexity) for queries with entries specified by b bits (integers range in $[-2^b, 2^b]$),

$$- \text{ (any } p) \ m = \Omega\left(\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon}\right)^p \frac{1}{b + \log(1/\epsilon)}\right)$$

- (RAM model) for $\epsilon \in (0, \log^{1/2-1/p}(1/\delta))$,

$$- \text{ (} p \in [1, 2] \text{) } m = \Omega\left(\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon}\right)^p + \frac{\log(1/\delta)}{\log \log(1/\delta)}\right)$$

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

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 general 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, \dots, 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 $\text{tr}(A_i)$ by t_i with $\mathbb{P}(|t_i - \text{tr}(A_i)| > \epsilon) \leq \delta$.

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 $[-2^b, 2^b]$),

$$- \text{ (} p \in [1, 2] \text{) } m = \Omega\left(\alpha m \left(\frac{1}{\epsilon^p(b + \log(1/\epsilon))} + \frac{\log(1/\delta)}{b + \log \log(1/\delta)}\right)\right)$$

$$- \text{ (} p \in [1, 2], \epsilon, \delta \in (0, 1/4) \text{) } m = \Omega\left(m \min\left(1, \frac{\alpha}{\epsilon}\right) \frac{\log(1/\delta)}{b + \log \log(1/\delta)}\right)$$

- (RAM model) for $\epsilon \in (0, \log^{1/2-1/p}(1/\delta))$,

$$- \text{ (} p \in [1, 2] \text{) } m = \Omega\left(\alpha m \left(\left(\frac{\sqrt{\log(1/\delta)}}{\epsilon}\right)^p\right)\right)$$

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 Hutchinson's estimator by sketching a low rank approximation of A and estimating the remaining component by Hutchinson's estimator, 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 $\text{tr}(A)$.

- 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** $t = \text{tr}(AQQ^T) + \frac{3}{m}\text{tr}(G^T A(I - QQ^T)G)$.
-

Remark 3.1.3. Here, we take the randomness of Gaussian [Jiang et al., 2021], while the original version 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 $(1 - \epsilon)\text{tr}(A) \leq t \leq (1 + \epsilon)\text{tr}(A)$.

Proof sketch for Theorem 3.1.6. First, note that $t = \text{tr}(AQQ^T) + H_{m/3}(A(I - QQ^T))$, 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 $|t - \text{tr}(A)| = |\text{tr}(\Delta) - H_{m/3}(\Delta)| \lesssim \sqrt{\frac{\log(1/\delta)}{m/3}} \|\Delta\|_F$ with probability (of G) at least $1 - \delta$.

Then, if $m = O\left(k + \sqrt{\log(1/\delta)}\right)$, by theorem 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 $\|\Delta\|_F \leq \|A - AQ_kQ_k^T\|_F \leq \sqrt{\frac{1+1/9}{1-1/9}} \|A - A_k\|_F \leq \sqrt{2} \|A - A_k\|_F$ and thus $\|\Delta\|_F \leq \sqrt{2} \|A - A_k\|_F$ 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 $|t - \text{tr}(A)| \lesssim \sqrt{\frac{\log(1/\delta)}{mk}} \text{tr}(A) \lesssim \epsilon \cdot \text{tr}(A)$ 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 $\text{tr}(A)$.

- 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** $t = \text{tr}(AR(S^T AR)^+ S^T A) + \frac{1}{c_3 m} \text{tr}(G^T (A - [AR(S^T AR)^+ S^T A])G)$ 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 $(1 - \epsilon)\text{tr}(A) \leq t \leq (1 + \epsilon)\text{tr}(A)$.

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

Proof sketch for Theorem 3.1.7. Similar to the adaptive one, here we divide A into $\tilde{A} := AR(S^T AR)^+ S^T A$ 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 $\|A - A_k\|_F = \min_X \|A - A_k X\|_F$. Consider regression sketching $\bar{X} = \arg \min_X \|R^T(A - A_k X)\|_F = (R^T A_k)^+ R^T A$, then if $c_2 m = O(k + \log(1/\delta))$, 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 $\|A - A_k \bar{X}\|_F = \|A - A_k (R^T A_k)^+ R^T A\|_F = \|A - AR(A_k R)^+ A_k\|_F \geq \|A - ARX^*\|_F$, where $X^* = \arg \min_X \|ARX - A\|_F$. Then another regression sketching yields that, if $c_1 m = O(c_2 m + \log(1/\delta))$, we have $O(1)\|A - ARX^*\|_F \geq \|A - AR\tilde{X}\|_F$, where $\tilde{X} := \arg \min_X \|S^T(ARX - A)\|_F = (S^T AR)^+ S^T A$.

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

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 algorithm *xiaobo*: (add ref) solves the problem 3.1.3 using at most*

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

(adaptive) queries of matrix-vector oracles.

Proof sketch for Theorem 3.1.8. \square

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]

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*

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

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 $R(N) := (x, y) : (x, y) \text{ can reach node } N \text{ always forms a square, i.e., } R(N) = A \times B \text{ for some } A \subseteq \mathcal{X} \text{ and } B \subseteq \mathcal{Y}$. 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 simplest among them.

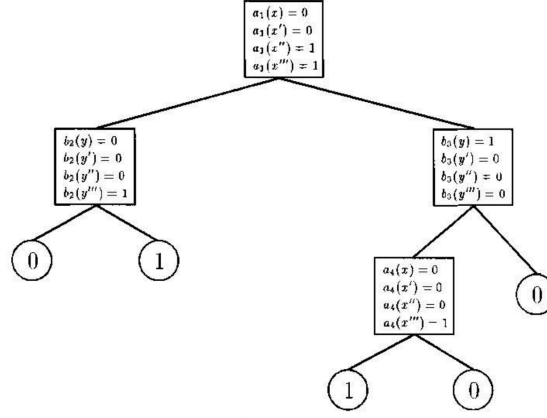


Figure 1.1: A protocol tree

	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 $\exists \epsilon, \delta \in (0, 1)$ such that μ satisfies*

$$\mu(R \cap f^{-1}(+1)) > \epsilon \mu(R \cap f^{-1}(-1))$$

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

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

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 $\{\pm 1\}^n \times \{\pm 1\}^n$ 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. \square

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

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

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

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 \square

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

$$1 \text{ if } s = t \quad \text{and} \quad -1 \text{ otherwise.}$$

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 $y_{i+1} = \dots = y_n = \perp$ for some unique i . The Gap-Equality problem asks Alice and Bob to return

$$1 \text{ if } s_i = t_i \quad \text{and} \quad -1 \text{ otherwise.}$$

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 $v, v' \stackrel{\text{dist}}{=} z$ 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 $1/3$. \square

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, \dots, q_m and oracles Aq_1, \dots, 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 $BQ(B \sim P \text{ or } Q \text{ respectively})$. 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 $l = n + (n-1) + \dots + (n-m+1)$. Observe that conditioned on a realization g we have

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

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 $\mathcal{E} = \left\{ \frac{\|g^T Q\|_2^2}{\|g\|_2^2} \leq \frac{1}{50C^2 n^3} \right\}$ and show that $\mathbb{P}(\mathcal{E}) \geq 10\delta$ if we only take $m = O\left(\frac{\log(1/\delta)}{\log \log(1/\delta)}\right)$ queries. Indeed, assume $m \leq n/2$

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

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

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

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 $d_{TV}(P, Q) \leq 1 - 2\delta$ instead?

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 measure 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 $\delta > 0$ such that for all large enough n , if A, B both have measure $\gamma(A), \gamma(B) \geq e^{-\delta n}$ then*

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

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*

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

Remark 5.2.1. *This lemma incorporates lemma 3.3 in [Sherstov, 2012](key of theorem 4.1.2) as a special 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 $\text{tr}(A)$ by*

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

where $G = [g_1, \dots, g_m] \in \mathbb{R}^{d \times m}$ 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 $H_\ell(A)$ 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 $\geq 1 - \delta$,*

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

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

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

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 $x \in \{-1, +1\}^n$,*

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

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*

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

where $c > 0$ is an absolute constant.

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

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

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

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

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

$$\begin{aligned} \mathbb{P}[\rho(x, V) > m + t] &\leq 2e^{-t^2/16}, \\ \mathbb{P}[\rho(x, V) \leq m - t] &\leq 2e^{-t^2/16}. \end{aligned}$$

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

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*

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

Theorem 5.2.5 (simple sketching is projection-cost-preserving [Musco and Musco, 2020]). *Let $S \in \mathbb{R}^{n \times m}$ be a i.i.d. standard Gaussian matrix. If $m \geq c(k + \log(1/\delta))/\epsilon^2$ 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. \square

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 $\|\tilde{A} - \tilde{P}\tilde{A}\|_F^2 \leq \gamma \min_{P \in \mathcal{T}} \|\tilde{A} - P\tilde{A}\|_F^2$, we have*

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

Proof for Theorem 5.2.6. Direct calculation. \square

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 $\text{rank}(A) = k$. Let $S \in \mathbb{R}^{r \times n}$ be a random matrix with i.i.d. Gaussian $\mathcal{N}(0, \frac{1}{r})$. Let $\tilde{X} = \arg \min_X \|S(AX - B)\|_F$ and $X^* = \arg \min_X \|AX - B\|_F$. Then, if $r = \Omega(k + \log(1/\delta))$, we have with probability at least $1 - \delta$*

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

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

Proof sketch for Theorem 5.2.7. \square

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 $1 - \delta \leq \|x\| \leq 1 + \delta$. 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 $\mathbb{E}_{P_n} \left[\left(\frac{dQ_n}{dP_n} \right)^2 \right] = \mathbb{E}_{Q_n} \left[\frac{dQ_n}{dP_n} \right]$

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

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

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

Finally, passing to the second moment, we compute

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

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

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

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

Proof sketch for Theorem 5.3.2. \square

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

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

Proof sketch for Theorem 5.3.3. \square

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, \dots, v_T , along with oracles $w_i = W v_i$. Then, conditioned on these observations, there exists a rotation matrix U , independent of w_i , such that

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

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 $\|A - A_k\|_F \leq \frac{\text{tr}(A)}{\sqrt{k}}$.*

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

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