Distributed Matrix Multiplication: Download Rate, Randomness and Privacy Trade-Offs

Amirhosein Morteza
Department of Computer Science & Engineering
University of Texas at Arlington
Arlington, TX 76019
amirhosein.morteza@uta.edu

Rémi A. Chou
Department of Computer Science & Engineering
University of Texas at Arlington
Arlington, TX 76019
remi.chou@uta.edu

2024 60th Annual Allerton Conference on Communication, Control, and Computing (Allerton) | 979-8-3315-4103-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/ALLERTON63246.2024.10735263

***Abstract*—We study the trade-off between communication rate and privacy for distributed batch matrix multiplication of two independent sequences of matrices $\mathbf{A}$ and $\mathbf{B}$ with uniformly distributed entries. In our setting, $\mathbf{B}$ is publicly accessible by all the servers while $\mathbf{A}$ must remain private. A user is interested in evaluating the product $\mathbf{AB}$ with the responses from the k fastest servers. For a given parameter $\alpha \in [0,1]$, our privacy constraint must ensure that any set of ℓ colluding servers cannot learn more than a fraction α of $\mathbf{A}$. Additionally, we study the trade-off between the amount of local randomness needed at the encoder and privacy, which to the best of our knowledge no previous work has characterized. Finally, we establish the optimal trade-offs when the matrices are square and identify a linear relationship between information leakage and communication rate.**

***Index Terms*—Secure Distributed Matrix Multiplication, Secret Sharing, Data Privacy, Distributed Computing.**

I. Introduction

THE task of multiparty computation with security guarantees, first explored in [1], [2], recently developed into outsourcing large-scale matrix multiplication tasks to distributed servers to speed up computation. It has applications in machine learning, signal processing, data encryption, and computational efficiency in cloud computing, e.g., [3]–[6]. For instance, privacy-preserving machine learning may involve training on encrypted data and protecting private information like health records. In several applications, only one matrix needs to be private. [7] and [8] explore secure outsourcing matrix computations in neural networks, requiring the privacy of input matrices ($\mathbf{A}$) while allowing computation results or model parameters ($\mathbf{B}$) to be openly handled. This setup exemplifies the principle of keeping sensitive data private while utilizing public model parameters.

The information-theoretic investigation of secure distributed matrix multiplication emerged in [9], where two matrices A and B are securely encoded and transmitted to N servers by the user who retrieves AB from the data downloaded. However, a scenario where a controlled amount of information leakage is permissible can help reduce communication complexity. To study the trade-off between privacy leakage and communication rate, we consider a setting with a newly defined privacy constraint that allows a controlled amount of leakage and design a coding scheme that meets such constraint.

In this paper, we study the problem of secure batch matrix multiplication for two sequences of matrices, $\mathbf{A}$ and $\mathbf{B}$, independently and uniformly distributed over a finite field. The user is interested in distributing the computation of the product $\mathbf{AB}$ over N servers. By downloading responses from the k fastest servers, the user can retrieve $\mathbf{AB}$ (Recoverability Constraint). The download rate is the ratio of the number of bits required to represent the computation result to the total number of bits that the servers must transmit to the user. Unlike previous studies, which considered perfect privacy, e.g., [6], [9], [10], in our setting, a controlled amount of information leakage is permissible, meaning that, for a given parameter $\alpha \in [0,1]$, no more than a fraction α of information about $\mathbf{A}$ can be learned by any set of ℓ colluding servers (Privacy Constraint). The capacity is defined as the supremum of the download rate. The capacity of this model has been characterized in [9], [10] when $\alpha = 0$, and its characterization when the matrices are non-uniform is an open problem [10]. Our main contributions are:

(i) Formalizing a new problem setting that enables the study of the trade-off between download rate and privacy leakage. Our results generalize [9, Theorem 1] and [10, Theorem 3] obtained when $\alpha = 0$.

(ii) Understanding the trade-off between privacy leakage and local randomness to efficiently use this costly resource at the encoder. Specifically, we determine bounds for the amount of randomness needed to meet the privacy constraint, which to the best of our knowledge, no previous works had investigated, even when $\alpha = 0$.

(iii) Characterizing the capacity for square matrices and showing a download rate gain of $\min\left(\frac{(k-\ell)\alpha}{k(1-\alpha)}, \frac{\ell}{k}\right)$ compared to the case $\alpha = 0$. We also identify a linear relationship between communication rate and privacy leakage by finding that the capacity is proportional to $\frac{1}{1-\alpha}$. Finally, we establish the optimal rate of local randomness needed at the encoder and show a gain of $\min\left(\frac{\alpha k}{k-\ell}, \frac{\ell}{k-\ell}\right)$ compared to the case $\alpha = 0$.

This work was supported in part by NSF grant CCF-2425371.

A. Related works

[9] is the first information-theoretic work that studies the capacity for secure distributed multiplication of two matrices A and B. The authors designed two models: a one-sided secure model where only matrix A is private, and a fully secured model where both matrices are private. Recent works, e.g., [5], [6], [9]–[11], aimed to optimize communication overheads in this problem. This problem is also explored in [11]–[13], [13]–[22] to investigate the use of coding techniques that reduce communication costs. Additionally, [23] and [24] considered the setting with distributed nodes where data does not originate at the user requesting the computation. These works focus on reducing the download and upload rate while preserving privacy.

Note that the above references only considered perfect security without any information leakage in their problem settings. Notably, [25] has explored the trade-off between privacy and sparsity in distributed computing by examining sparse secret-sharing schemes, where increased sparsity results in weaker privacy.

Compared to previous works, our study demonstrates how relaxing privacy constraints can enhance communication efficiency in distributed matrix multiplication.

B. Main differences with previous works

The recent models in [10], [6], and [13], use matrix partitioning techniques introduced in [26]. These methods consider no privacy leakage, ensuring that any set of colluding servers cannot obtain any information about private matrices. To the best of our knowledge, no previous work characterizes the capacity for secure distributed matrix multiplication when information leakage is allowed. In this study, we define a new problem setting that incorporates a privacy constraint with an information leakage parameter α, which could not be addressed with previous techniques, as detailed next. With such a privacy constraint, we obtain bounds on the capacity that generalize previously found bounds in [9, Theorem 1] and [10, Theorem 3]. Another contrast between this work and previous studies is that we bound the optimal rate of local randomness needed at the encoder to satisfy the privacy constraint, making this the first study to explore such bounds.

In our setting, a significant challenge lies in satisfying the new privacy constraint, preventing any set of $\ell < k$ colluding servers to learn more than a fraction $\alpha \in [0, 1]$ of information about $\mathbf{A}$. We cannot rely on traditional secret sharing [27], [28], which does not allow any information leakage. We modify ramp secret-sharing schemes [29] to integrate the matrix multiplication task. Inspired by recent studies on the trade-offs between privacy and communication rate [30], as well as storage considerations [31], we combine two ramp coding schemes—a strategy not previously investigated—to allow a controlled privacy leakage.

C. Paper organization

The remainder of the paper is organized as follows. We define the problem in Section II and present our main results in Section III. We discuss our converse and achievability in Sections IV and V, respectively.

II. Problem Statement

Notation: Let $\mathbb{F}_q$ be a finite field characterized by a large prime number q. Let $\mathbb{Q}$, $\mathbb{N}$, and $\mathbb{R}$ be the set of rational, natural, and real numbers, respectively. For any $a, b \in \mathbb{N}$, define $[a] \triangleq [1, a] \cap \mathbb{N}$, $[a : b] \triangleq [a, b] \cap \mathbb{N}$, and $[a : b) \triangleq [a, b-1] \cap \mathbb{N}$. Sets are represented by calligraphic letters, and sequences of matrices are represented by bold uppercase letters. Let $[a]^{=b} \triangleq \{\mathcal{I} \subseteq [a] : |\mathcal{I}| = b\}$ be the set of all the subsets of $[a]$ that have cardinality b; $[a]^{\leq b} \triangleq \{\mathcal{I} \subseteq [a] : |\mathcal{I}| \leq b\}$ be the set of all the subsets of $[a]$ that have a cardinality less than or equal to b. Logarithms are defined with base q. Also, define $[a]^+ \triangleq \max\{0, a\}$.

Definition 1. *Let $N, r \in \mathbb{N}$, and $k \in [N]$, an (N, k, r)-coding scheme consists of*

- *$N \geq 2$ servers;*
- *Two sequences of matrices, $\mathbf{A} \triangleq (A_s)_{s\in[m]}$ and $\mathbf{B} \triangleq (B_s)_{s\in[m]}$, where m is a large integer. For any $s \in [m]$, the matrices A_s and B_s are assumed to be independent and uniformly distributed over $\mathbb{F}_q^{C\times D}$ and $\mathbb{F}_q^{D\times E}$, respectively. $\mathbf{B}$ is public while $\mathbf{A}$ is private and accessible only by the user;*
- *Local randomness in the form of a uniform random variable R which is distributed over $\mathbb{F}_q^r$ and independent of $(\mathbf{A}, \mathbf{B})$;*
- *N encoding functions $f_i : (\mathbf{A}, R) \mapsto \tilde{\mathbf{A}}_i$, $i \in [N]$, such that $\mathbf{A}$ can be recovered from encoded matrices $\tilde{\mathbf{A}}_\mathcal{I} \triangleq (\tilde{\mathbf{A}}_i)_{i\in\mathcal{I}}$, $\mathcal{I} \in [N]^{\geq k}$, i.e.,*

$$H(\mathbf{A}|\tilde{\mathbf{A}}_\mathcal{I}) = 0; \quad (1)$$

- *N processing functions $h_i : (\tilde{\mathbf{A}}_i, \mathbf{B}) \mapsto Z_i$, $i \in [N]$;*
- *A decoding function d taking $Z_\mathcal{I} \triangleq (Z_i)_{i\in\mathcal{I}}$, $\mathcal{I} \subseteq [N]$, and returning an estimate of $\mathbf{AB} \triangleq (A_sB_s)_{s\in[m]}$;*

and operates as follows:

- *For all $i \in [N]$, the user sends the encoded matrices $\tilde{\mathbf{A}}_i \triangleq f_i(\mathbf{A}, R)$ to Server i over a private channel;*
- *For all $i \in [N]$, Server i generates a response $Z_i \triangleq h_i(\tilde{\mathbf{A}}_i, \mathbf{B})$;*
- *The user computes an estimate of $\mathbf{AB}$ as $d(Z_\mathcal{I})$, where $Z_\mathcal{I}$ is a sequence of received responses from the k fastest servers.*

Definition 2. *For any $\ell \in [0 : k)$ and $\alpha \in [0, 1) \cap \mathbb{Q}$, an (N, k, r)-coding scheme is (ℓ, α)-private if*

$$\max_{\mathcal{I}\in[N]^{\geq k}} H(\mathbf{AB}|Z_\mathcal{I}) = 0, \qquad (Recoverability) \quad (2)$$

$$\max_{\mathcal{L}\in[N]^{\leq \ell}} \frac{I(\mathbf{A}; \tilde{\mathbf{A}}_\mathcal{L})}{H(\mathbf{A})} \leq \alpha. \qquad (Privacy) \quad (3)$$

An achievable rate Λ_k for an (N, k, r)-coding scheme that satisfies (2) *and* (3) *is determined by the ratio of the desired*

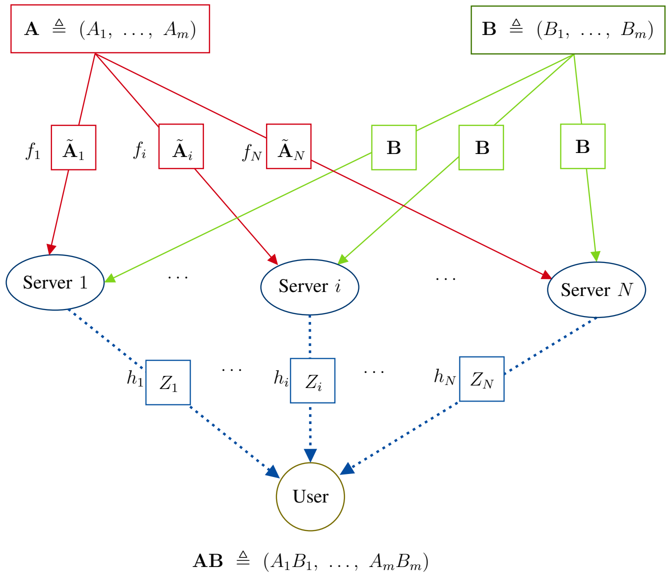

Figure 1. The matrices in $\mathbf{A}$ are first encoded by the user using the functions f_i. These encoded matrices, denoted as $\tilde{\mathbf{A}}_i$, are then distributed to each Server $i \in [N]$. Server i processes $\tilde{\mathbf{A}}_i$ and $\mathbf{B}$ to compute a response Z_i using the function h_i. Once the computations are complete, the user collects the responses from the k fastest servers and reconstructs the product $\mathbf{AB}$.

information to the total download from the k fastest servers

$$\Lambda_k \triangleq \frac{H(\mathbf{AB}|\mathbf{B})}{\max_{\mathcal{I}\in[N]^{=k}} \sum_{i\in\mathcal{I}} H(Z_i)}. \tag{4}$$

The capacity $C_{(\ell,\alpha,k)}$ is the supremum of all achievable rates. Additionally, we define the optimal rate of local randomness as $R_{(\ell,\alpha,k)} \triangleq \min\{r \in \mathbb{N} : \exists(\ell,\alpha)$-private (N,k,r)-coding scheme$\}/H(\mathbf{AB}|\mathbf{B})$.

Equation (2) means that the responses from any subset of k or more servers are sufficient to reconstruct the product $\mathbf{AB}$. Equation (3) means that ℓ colluding servers cannot learn more than a fraction α of $\mathbf{A}$ from the encoded matrices $\tilde{\mathbf{A}}_{\mathcal{L}}$. Note that α is chosen as a rational number, which is not a restrictive assumption because by density of $\mathbb{Q}$ in $\mathbb{R}$, for any $\beta \in [0,1]$, $\epsilon > 0$, there exists $\alpha \in [0,1) \cap \mathbb{Q}$ such that $|\alpha - \beta| < \epsilon$.

In Equation (4), we consider the maximum over any subset of k servers in the denominator to account for the worst-case scenario.

Fig. 1 illustrates our setting.

III. Main Results

The following theorem establishes upper and lower bounds on the capacity.

Theorem 1 (Communication Rate)**.** *For any $\alpha \in [0,1) \cap \mathbb{Q}$ and $\ell \in [0:k)$, the capacity $C_{(\ell,\alpha,k)}$ satisfies*

$$C_{(\ell,\alpha,k)} \leq \begin{cases} \min\left(\frac{k-\ell}{k\left(1-\alpha\max\left(1,\frac{D}{E}\right)\right)}, 1\right) & \alpha < \frac{1}{\max\left(1,\frac{D}{E}\right)} \\ 1 & \alpha \geq \frac{1}{\max\left(1,\frac{D}{E}\right)} \end{cases}, \tag{5}$$

$$C_{(\ell,\alpha,k)} \geq \begin{cases} \frac{k-\ell}{k(1-\alpha)} \min\left(1, \frac{D}{E}\right) & \alpha < \frac{\ell}{k} \\ \min\left(1, \frac{D}{E}\right) & \alpha \geq \frac{\ell}{k} \end{cases}. \tag{6}$$

Proof. We prove the converse and achievability in Sections IV and V, respectively. □

Theorem 2 (Local Randomness)**.** *For any $\alpha \in [0,1) \cap \mathbb{Q}$ and $\ell \in [1:k)$, the optimal rate of local randomness necessary at the encoder satisfies*

$$R_{(\ell,\alpha,k)} \leq \begin{cases} \frac{\ell-\alpha k}{k-\ell} \max\left(1, \frac{D}{E}\right) & \alpha < \frac{\ell}{k} \\ 0 & \alpha \geq \frac{\ell}{k} \end{cases}, \tag{7}$$

$$R_{(\ell,\alpha,k)} \geq \frac{\left[\ell - \alpha k \max\left(1, \frac{D}{E}\right)\right]^+}{k-\ell}. \tag{8}$$

Proof. We prove the converse and achievability in Sections IV-B and V-B2, respectively. □

The bounds established in the previous theorems match when the matrices are square.

Theorem 3 (Optimality Results)**.** *If $\mathbf{A}$ and $\mathbf{B}$ are two sequences of independent and square matrices with uniformly distributed entries, then the capacity is*

$$C_{(\ell,\alpha,k)} = \min\left(\frac{k-\ell}{k(1-\alpha)}, 1\right) = \begin{cases} \frac{k-\ell}{k(1-\alpha)} & \alpha < \frac{\ell}{k} \\ 1 & \alpha \geq \frac{\ell}{k} \end{cases},$$

and the optimal rate of local randomness is

$$R_{(\ell,\alpha,k)} = \frac{[\ell - \alpha k]^+}{k-\ell} = \begin{cases} \frac{\ell-\alpha k}{k-\ell} & \alpha < \frac{\ell}{k} \\ 0 & \alpha \geq \frac{\ell}{k} \end{cases}.$$

Proof. One can deduce these results from the bounds in Theorems 1 and 2 with $D = E$. □

IV. Converse Proof

Define $\mathcal{L} \subseteq [N]$ such that $|\mathcal{L}| = \ell$.

A. Rate

Initially, consider $\alpha < \frac{1}{\max\left(1,\frac{D}{E}\right)}$. For all $\mathcal{I} \in [N]^{=k}$ such that $\mathcal{L} \subseteq \mathcal{I}$, we have

$$\begin{aligned} H(\mathbf{AB}|\mathbf{B}) &= H(\mathbf{AB}|\mathbf{B}) - H(\mathbf{AB}|Z_{\mathcal{I}},\mathbf{B}) + H(\mathbf{AB}|Z_{\mathcal{I}},\mathbf{B}) \\ &\overset{(a)}{=} I(Z_{\mathcal{I}};\mathbf{AB}|\mathbf{B}) \\ &= H(Z_{\mathcal{I}}|\mathbf{B}) - H(Z_{\mathcal{I}}|\mathbf{AB},\mathbf{B}) \\ &\overset{(b)}{\leq} H(Z_{\mathcal{I}}|\mathbf{B}) - H(Z_{\mathcal{L}}|\mathbf{AB},\mathbf{B}) & (9) \\ &\overset{(c)}{\leq} H(Z_{\mathcal{I}}|\mathbf{B}) - H(Z_{\mathcal{L}}|\mathbf{B}) + \alpha H(\mathbf{A}), & (10) \end{aligned}$$

where

(a) holds by (2);
(b) holds because $\mathcal{L} \subseteq \mathcal{I}$;

(c) holds because $H(Z_{\mathcal{L}}|\mathbf{AB},\mathbf{B}) \geq H(Z_{\mathcal{L}}|\mathbf{B}) - \alpha H(\mathbf{A})$. The proof is omitted due to space constraints.

Then, we have

$$
\begin{aligned}
H(\mathbf{AB}|\mathbf{B}) &\overset{\text{(a)}}{\leq} H(Z_{\mathcal{I}}|\mathbf{B}) - \ell \frac{1}{\binom{k}{\ell}} \sum_{\mathcal{L}\in\mathcal{I}^{=\ell}} \frac{H(Z_{\mathcal{L}}|\mathbf{B})}{\ell} + \alpha H(\mathbf{A}) \\
&\overset{\text{(b)}}{\leq} H(Z_{\mathcal{I}}|\mathbf{B}) - \ell \frac{H(Z_{\mathcal{I}}|\mathbf{B})}{k} + \alpha H(\mathbf{A}) \\
&= \left(1 - \frac{\ell}{k}\right) H(Z_{\mathcal{I}}|\mathbf{B}) + \alpha H(\mathbf{A}) \\
&\leq \left(1 - \frac{\ell}{k}\right) \sum_{i\in\mathcal{I}} H(Z_i) + \alpha H(\mathbf{A}), \qquad (11)
\end{aligned}
$$

where

(a) holds by averaging (10) over all possible subsets $\mathcal{L}$ of servers of size ℓ in $\mathcal{I}$;

(b) holds by Han's inequality [32, section 17.6].

Then, from (11) we have

$$
\begin{aligned}
\Lambda_k &= \frac{H(\mathbf{AB}|\mathbf{B})}{\max_{\mathcal{I}\in[N]^{=k}} \sum_{i\in\mathcal{I}} H(Z_i)} \\
&\leq \frac{1 - \frac{\ell}{k}}{1 - \alpha \frac{H(\mathbf{A})}{H(\mathbf{AB}|\mathbf{B})}}. \qquad (12)
\end{aligned}
$$

Given that the matrices in $\mathbf{A}$ and $\mathbf{B}$ are independent and uniformly distributed over $\mathbb{F}_q^{C\times D}$ and $\mathbb{F}_q^{D\times E}$, using [10, Lemma 2], we have

$$
q \to \infty \Rightarrow H(\mathbf{AB}|\mathbf{B}) = m \times \min(CD, CE). \qquad (13)
$$

Also, for any $\alpha \in [0,1)$, we have from (9)

$$
\begin{aligned}
H(\mathbf{AB}|\mathbf{B}) &\leq H(Z_{\mathcal{I}}|\mathbf{B}) - H(Z_{\mathcal{L}}|\mathbf{AB},\mathbf{B}) \\
&\overset{\text{(a)}}{=} H(Z_{\mathcal{I}}|\mathbf{B}) \\
&\overset{\text{(b)}}{\leq} \sum_{i\in\mathcal{I}} H(Z_i), \qquad (14)
\end{aligned}
$$

where

(a) holds by (9) with $|\mathcal{L}| = 0$;

(b) holds because conditioning reduces entropy.

Then, from (14), we have

$$
\begin{aligned}
\Lambda_k &= \frac{H(\mathbf{AB}|\mathbf{B})}{\max_{\mathcal{I}\in[N]^{=k}} \sum_{i\in\mathcal{I}} H(Z_i)} \\
&\leq \frac{\sum_{i\in\mathcal{I}} H(Z_i)}{\max_{\mathcal{I}\in[N]^{=k}} \sum_{i\in\mathcal{I}} H(Z_i)} \\
&\leq 1. \qquad (15)
\end{aligned}
$$

Finally, we obtain (5) from (12), (13) and (15).

B. Local Randomness

Consider $\alpha \in [0,1] \cap \mathbb{Q}$ and $\ell \in [1:k)$. Let $\mathcal{V} \subseteq \mathcal{L}$ define $v \triangleq |\mathcal{V}|$, $\mathcal{V}_0 \triangleq \emptyset$, and for $j \in \mathcal{L}$, $\mathcal{V}_j \triangleq \mathcal{V}_{j-1} \cup \{i^*(\mathcal{V})\}$. Then, we have

$$
\begin{aligned}
&\frac{\ell - k\alpha\left(\frac{H(\mathbf{A})}{H(\mathbf{AB}|\mathbf{B})}\right)}{k-\ell} H(\mathbf{AB}|\mathbf{B}) \\
&= -\alpha H(\mathbf{A}) + \ell \frac{1 - \alpha\frac{H(\mathbf{A})}{H(\mathbf{AB}|\mathbf{B})}}{k-\ell} H(\mathbf{AB}|\mathbf{B}) \\
&\overset{(a)}{\leq} -\alpha H(\mathbf{A}) + \sum_{i=0}^{\ell-1} \left[H(\tilde{\mathbf{A}}_{[N]}|\tilde{\mathbf{A}}_{\mathcal{V}_i}) - H(\tilde{\mathbf{A}}_{[N]}|\tilde{\mathbf{A}}_{\mathcal{V}_{i+1}}) \right] \\
&= -\alpha H(\mathbf{A}) + H(\tilde{\mathbf{A}}_{[N]}) - H(\tilde{\mathbf{A}}_{[N]}|\tilde{\mathbf{A}}_{\mathcal{V}_\ell}) \\
&\overset{(b)}{\leq} -\alpha H(\mathbf{A}) + H(\mathbf{A}, R) - H(\tilde{\mathbf{A}}_{[N]}|\tilde{\mathbf{A}}_{\mathcal{V}_\ell}) \\
&\overset{(c)}{=} (1-\alpha) H(\mathbf{A}) + H(R) - H(\tilde{\mathbf{A}}_{[N]}|\tilde{\mathbf{A}}_{\mathcal{V}_\ell}) \\
&\overset{(d)}{=} (1-\alpha) H(\mathbf{A}) + H(R) - H(\mathbf{A}\tilde{\mathbf{A}}_{[N]}|\tilde{\mathbf{A}}_{\mathcal{V}_\ell}) \\
&\leq (1-\alpha) H(\mathbf{A}) + H(R) - H(\mathbf{A}|\tilde{\mathbf{A}}_{\mathcal{V}_\ell}) \\
&\overset{(e)}{\leq} H(R), \qquad (16)
\end{aligned}
$$

where

(a) holds by the definition of $\mathcal{V}_j$, $j \in \mathcal{L}$, and applying ℓ times the following inequality

$$
\begin{aligned}
&\frac{1 - \alpha\frac{H(\mathbf{A})}{H(\mathbf{AB}|\mathbf{B})}}{k-\ell} H(\mathbf{AB}|\mathbf{B}) \\
&\leq H(\tilde{\mathbf{A}}_{[N]}|\tilde{\mathbf{A}}_{\mathcal{V}}) - H(\tilde{\mathbf{A}}_{[N]}|\tilde{\mathbf{A}}_{\mathcal{V}\cup\{i^*(\mathcal{V})\}}), \qquad (17)
\end{aligned}
$$

where $i^*(\mathcal{V}) \in \arg\max_{i\in[N]\setminus\mathcal{V}} H(\tilde{\mathbf{A}}_i|\tilde{\mathbf{A}}_{\mathcal{V}})$ and the proof for (17) is omitted due to space constraints;

(b) holds because $\tilde{\mathbf{A}}_{[N]}$ is a deterministic function of $(\mathbf{A}, R)$;

(c) holds by independence between $\mathbf{A}$ and R;

(d) holds because, by (1), we have

$$
\begin{aligned}
H(\mathbf{A}\tilde{\mathbf{A}}_{[N]}|\tilde{\mathbf{A}}_{\mathcal{V}_\ell}) &= H(\mathbf{A}|\tilde{\mathbf{A}}_{[N]}) + H(\tilde{\mathbf{A}}_{[N]}|\tilde{\mathbf{A}}_{\mathcal{V}_\ell}) \\
&= H(\tilde{\mathbf{A}}_{[N]}|\tilde{\mathbf{A}}_{\mathcal{V}_\ell});
\end{aligned}
$$

(e) holds because $-H(\mathbf{A}|\tilde{\mathbf{A}}_{\mathcal{V}_\ell}) \leq -(1-\alpha)H(\mathbf{A})$ by (3).

Using (16) and (13), we have

$$
\frac{\ell - k\alpha\left(\max\left(1, \frac{D}{E}\right)\right)}{k-\ell} H(\mathbf{AB}|\mathbf{B}) \leq H(R). \qquad (18)
$$

Finally, (8) holds by (18), and since $H(R) \geq 0$.

V. Achievability Proof

For the achievability, the idea is to design a coding scheme with the following leakage symmetry condition

$$
\begin{aligned}
&\forall t \in [N], \exists\, E_t \in [0,1], \forall\, \mathcal{I} \subseteq [N], \\
&|\mathcal{I}| = t \Rightarrow \frac{I(\mathbf{A}; \tilde{\mathbf{A}}_{\mathcal{I}})}{H(\mathbf{A})} = E_t, \qquad (19)
\end{aligned}
$$

which means that the leakage of any set of encoded matrices $\tilde{\mathbf{A}}_{\mathcal{I}} \triangleq (\tilde{\mathbf{A}}_i)_{i\in\mathcal{I}}$ only depends on the cardinality of $\mathcal{I}$. Consequently, the amount of information leakage of $\mathbf{A}$ can be fully described by the function

$$
g : [N] \to [0,1],\ t \mapsto E_t.
$$

The recoverability and privacy constraints in Equations (2) and (3) impose the following constraints on g:

$$
\forall t \in [\ell],\ g(t) \leq \alpha, \qquad \forall t \in [k:N],\ g(t) = 1.
$$

Consider two cases: $\alpha \geq \frac{\ell}{k}$ and $\alpha < \frac{\ell}{k}$. The case $\alpha \geq \frac{\ell}{k}$ will be handled with a modified ramp secret-sharing scheme with g defined as

$$g : i \mapsto \begin{cases} \frac{i}{k}, & i \in [0:k) \\ 1, & i \in [k:N] \end{cases}. \tag{20}$$

In the case $\alpha < \frac{\ell}{k}$, we define g as

$$g = g_1 + g_2, \tag{21}$$

with

$$g_1 : i \mapsto \begin{cases} \frac{\alpha}{\ell} i, & i \in [0:k) \\ \frac{\alpha}{\ell} k, & i \in [k:N] \end{cases}, \tag{22}$$

$$g_2 : i \mapsto \begin{cases} 0, & i \in [0:\ell] \\ \frac{1-\alpha}{k-\ell}(i-\ell) + \alpha - \frac{\alpha}{\ell} i, & i \in (\ell: \ k) \\ 1 - \frac{\alpha}{\ell} k, & i \in [k:N] \end{cases}, \tag{23}$$

and construct a coding scheme by combining two modified ramp secret-sharing schemes with normalized access functions g_1 and g_2.

In Section V-A, we present our coding scheme. Then, in Section V-B, we analyze our coding scheme and show that it satisfies our setting constraints, (2) and (3).

A. *Coding Scheme*

1) Case 1: $\alpha \geq \frac{\ell}{k}$. We divide the sequence of m matrices in $\mathbf{A}$ and $\mathbf{B}$ into blocks of k matrices. Given $k < m$, there are $\lceil \frac{m}{k} \rceil$ blocks. For any $b \in \lceil \frac{m}{k} \rceil$, define

$$\mathcal{S}_b \triangleq [(b-1)k + 1 : bk].$$

Consider k distinct non-zero constants $x_i \in \mathbb{F}_q, i \in [k]$. For any $b \in \lceil \frac{m}{k} \rceil$, $i \in [N]$, define

$$\forall s \in \mathcal{S}_b, \tilde{A}_i^s \triangleq x_i^{s-1} A_s.$$

The response for each block from Server $i \in [N]$ is

$$\begin{aligned} \forall b \in \left\lceil \frac{m}{k} \right\rceil, Z_i^b &\triangleq \sum_{s \in \mathcal{S}_b} \tilde{A}_i^s B_s \\ &= \sum_{s \in \mathcal{S}_b} x_i^{s-1} A_s B_s. \end{aligned} \tag{24}$$

The total response from Server $i \in [N]$ is

$$Z_i \triangleq (Z_i^b)_{b \in \lceil \frac{m}{k} \rceil}. \tag{25}$$

Upon receiving responses from k servers, the user can decode the product matrices $(A_s \times B_s)_{s \in \mathcal{S}_b}$ for each block $\mathcal{S}_b$ as follows. Define

$$M_b \triangleq \begin{bmatrix} x_1^{(b-1)k} & x_1^{(b-1)k+1} & \dots & x_1^{bk-1} \\ x_2^{(b-1)k} & x_2^{(b-1)k+1} & \dots & x_2^{bk-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_k^{(b-1)k} & x_k^{(b-1)k+1} & \dots & x_k^{bk-1} \end{bmatrix} \in \mathbb{F}_q^{k \times k},$$

$$\forall b \in \lceil \tfrac{m}{k} \rceil, \begin{bmatrix} Z_1^b \\ Z_2^b \\ \vdots \\ Z_k^b \end{bmatrix} = M_b \begin{bmatrix} A_{(b-1)k+1} B_{(b-1)k+1} \\ A_{(b-1)k+2} B_{(b-1)k+2} \\ \vdots \\ A_{bk} B_{bk} \end{bmatrix}. \tag{26}$$

Equation (26) has a unique solution because M_b is invertible, since its determinant is a minor of a Vandermonde Matrix, for which each row i can be factored by $x_i^{(b-1)k}$, and $(x_i)_{i \in [k]}$ are non-zero and distinct.

2) Case 2: $\alpha < \frac{\ell}{k}$. Define the following index sets

$$\mathcal{P} \triangleq [p], p = \left\lceil \alpha \frac{k}{\ell} m \right\rceil, \tag{27}$$

$$\overline{\mathcal{P}} \triangleq [m] \setminus \mathcal{P}, |\overline{\mathcal{P}}| = m - p. \tag{28}$$

We partition $\mathbf{A}$ into two sub-sequences $A^{\mathcal{P}} \triangleq (A_j)_{j \in \mathcal{P}}$ and $A^{\overline{\mathcal{P}}} \triangleq (A_j)_{j \in \overline{\mathcal{P}}}$, then proceed as follows.

1) Break down the sequences $A^{\mathcal{P}}$ and $B^{\mathcal{P}}$ into blocks of k matrices. The encoding is the same as (24) and we have a matrix equation similar to (26). For $b \in \lceil \frac{p}{k} \rceil$, the response from Server $i \in [k]$ is

$$Z_i^{\mathcal{P}} \triangleq (Z_i^b)_{b \in \lceil \frac{p}{k} \rceil}. \tag{29}$$

2) Break down the sequence of matrices in $A^{\overline{\mathcal{P}}}$ and $B^{\overline{\mathcal{P}}}$ into blocks of $k - \ell$ matrices. For any $b \in \left\lceil \frac{m-p}{k-\ell} \right\rceil$, define

$$\mathcal{S}_b \triangleq [(b-1)(k-\ell) + 1 : b(k-\ell)]. \tag{30}$$

For any $s \in S_b$, let $R_s \triangleq (R_{(s,r)})_{r \in [\ell]}$ be uniformly distributed random matrices over $\mathbb{F}_q^{C \times D}$. Consider $k - \ell$ distinct non-zero constants $x_i \in \mathbb{F}_q, i \in [k-\ell]$, then for any $i \in [N]$, define

$$\forall s \in \mathcal{S}_b, \tilde{A}_i^s \triangleq x_i^{s-1} A_s + \sum_{r \in [\ell]} {x_i}^{r+(k-\ell)-1} R_{(s,r)}. \tag{31}$$

For $b \in \left\lceil \frac{m-p}{k-\ell} \right\rceil$, the response from Server $i \in [k]$ is

$$\begin{aligned} &Z_i^b \\ &\triangleq \sum_{s \in \mathcal{S}_b} \tilde{A}_i^s B_s \\ &= \sum_{s \in \mathcal{S}_b} (x_i^{s-1} A_s B_s) + \sum_{s \in \mathcal{S}_b} \sum_{r \in [\ell]} {x_i}^{r+(k-\ell)-1} R_{(s,r)} B_s. \end{aligned} \tag{32}$$

The user downloads Z_i^b and upon receiving k answers from the servers, recovers $(A_s \times B_s)_{s \in \mathcal{S}_b}$ from a matrix equation similar to (26) but with the following coefficient matrix,

$$M_b^{\overline{\mathcal{P}}} \triangleq \begin{bmatrix} x_1^{(b-1)(k-\ell)} & \dots & x_1^{b(k-\ell)-1} & \dots & {x_1}^{bk-1} \\ x_2^{(b-1)(k-\ell)} & \dots & x_2^{b(k-\ell)-1} & \dots & {x_2}^{bk-1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{k-\ell}^{(b-1)(k-\ell)} & \dots & x_{k-\ell}^{b(k-\ell)-1} & \dots & {x_{k-\ell}}^{bk-1} \\ x_{k-\ell+1}^{(b-1)(k-\ell)} & \dots & x_{k-\ell+1}^{b(k-\ell)-1} & \dots & {x_{k-\ell+1}}^{bk-1} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_k^{(b-1)(k-\ell)} & \dots & x_k^{b(k-\ell)-1} & \dots & x_k^{bk-1} \end{bmatrix} \in \mathbb{F}_q^{k\times k}.$$

The determinant of $M_b^{\overline{\mathcal{P}}}$ is also a minor of a Vandermonde matrix, for which each row i can be factored by $x_i^{(b-1)(k-\ell)}$. The response from Server $i \in [k]$ is

$$Z_i^{\overline{\mathcal{P}}} \triangleq (Z_i^b)_{b\in\lceil\frac{m-p}{k-\ell}\rceil}. \tag{33}$$

Finally, from (29) and (33) the total response Z_i from Server $i \in [N]$ is

$$Z_i \triangleq (Z_i^{\mathcal{P}}, Z_i^{\overline{\mathcal{P}}}). \tag{34}$$

B. Analysis of the Coding Scheme

1) Rate: For the case $\alpha \geq \frac{\ell}{k}$, we have

$$\begin{aligned}\Lambda_k &= \frac{H(\mathbf{AB}|\mathbf{B})}{\max_{\mathcal{I}\in[N]^{=k}}\sum_{i\in\mathcal{I}} H(Z_i)} \\ &\geq \frac{H(\mathbf{AB}|\mathbf{B})}{\lceil\frac{m}{k}\rceil kCE} \\ &= \frac{m\left(\min\left(\frac{D}{E},1\right)\right)}{k\lceil\frac{m}{k}\rceil} \\ &\xrightarrow{m\to\infty} \min\left(\frac{D}{E},1\right),\end{aligned}$$

where the inequality holds because for the entropy of the response Z_i, $i \in [N]$, defined in (25), we have $H(Z_i) \leq \lceil\frac{m}{k}\rceil CE$. The proof is omitted due to space constraints. Then, by summing up the entropy of the responses over k servers, we have

$$\max_{\mathcal{I}\in[N]^{=k}} \sum_{i\in\mathcal{I}} H(Z_i) \leq k\left\lceil\frac{m}{k}\right\rceil CE;$$

the second equality holds by (13).

For the case $\alpha < \frac{\ell}{k}$, from (34), we have

$$\forall i \in [N],\ H(Z_i) = H(Z_i^{\mathcal{P}}) + H(Z_i^{\overline{\mathcal{P}}}). \tag{35}$$

Then, we have

$$\begin{aligned}H(Z_i^{\mathcal{P}}) &\overset{\text{(a)}}{=} H\left((Z_i^b)_{b\in\lceil\frac{p}{k}\rceil}\right) \\ &\overset{\text{(b)}}{=} \left\lceil\frac{p}{k}\right\rceil H(Z_i^b) \\ &\overset{\text{(c)}}{\leq} \left\lceil\frac{p}{k}\right\rceil CE,\end{aligned} \tag{36}$$

where

(a) holds by (29);
(b) holds by independence of the blocks;
(c) holds because for any $s \in \mathcal{S}_b$, $\tilde{A}_i^s B_s \in \mathbb{F}_q^{C\times E}$.

Similarly, for matrices in blocks $\mathcal{S}_b$ defined in (30), we have

$$H(Z_i^{\overline{\mathcal{P}}}) \leq \left\lceil\frac{m-p}{k-\ell}\right\rceil CE. \tag{37}$$

Then, from (35), (36), and (37), we have

$$H(Z_i) \leq \left\lceil\frac{p}{k}\right\rceil CE + \left\lceil\frac{m-p}{k-\ell}\right\rceil CE. \tag{38}$$

For the Communication rate, we have

$$\begin{aligned}\Lambda_k &= \frac{H(\mathbf{AB}|\mathbf{B})}{\max_{\mathcal{I}\in[N]^{=k}}\sum_{i\in\mathcal{I}} H(Z_i)} \\ &\geq \frac{H(\mathbf{AB}|\mathbf{B})}{k\left(\lceil\frac{p}{k}\rceil + \left\lceil\frac{m-p}{k-\ell}\right\rceil\right)CE} \\ &= \frac{m\left(\min(1,\frac{D}{E})\right)}{k\left(\left\lceil\frac{\lceil\alpha\frac{k}{\ell}m\rceil}{k}\right\rceil + \left\lceil\frac{m-\lceil\alpha\frac{k}{\ell}m\rceil}{k-\ell}\right\rceil\right)} \\ &\xrightarrow{m\to\infty} \frac{\min\left(1,\frac{D}{E}\right)}{k(\frac{1-\alpha}{k-\ell})},\end{aligned}$$

where the first inequality holds by summing up the entropy of the responses over k servers and using (38), and the second equality holds by (13), (27) and (28).

2) Local Randomness: When $\alpha \geq \frac{\ell}{k}$, no randomness was used in the coding scheme, hence $R_{(\ell,\alpha,k)} = 0$. However, when $\alpha < \frac{\ell}{k}$, we use randomness in (32). For any $s \in \mathcal{S}_b$, we have

$$H(R_s) = \ell CD. \tag{39}$$

The proof for (39) is omitted due to space constraints. Then, the rate of local randomness is

$$\begin{aligned}\frac{R_{(\ell,\alpha,k)}}{H(\mathbf{AB}|\mathbf{B})} &= \frac{\left\lceil\frac{m-p}{k-\ell}\right\rceil H\left((R_s)_{s\in\mathcal{S}_b}\right)}{H(\mathbf{AB}|\mathbf{B})} \\ &\leq \frac{\ell\left\lceil\frac{m-\lceil\alpha\frac{k}{\ell}m\rceil}{k-\ell}\right\rceil}{m}\max(1,\frac{D}{E}) \\ &\xrightarrow{m\to\infty} \frac{\ell-\alpha k}{k-\ell}\max(1,\frac{D}{E}),\end{aligned}$$

where the equality holds because there are $\left\lceil\frac{m-p}{k-\ell}\right\rceil$ blocks and within each block, the R_s, $s \in S_b$ are jointly independent; the inequality holds by (28), (13), and (39).

3) Recoverability Constraint: For each block S_b, the user can recover $(A_sB_s)_{s\in S_b}$ because all the matrices M_b and $M_b^{\overline{\mathcal{P}}}$ are invertible, as explained in Sections V-A1 and V-A2, respectively. Hence, the coding scheme designed in Section V-A, satisfies (2).

4) Privacy Constraint: For any $\mathcal{L} \in [N]^{\leq \ell}$, we have

$$\begin{aligned} I(\mathbf{A}; \tilde{\mathbf{A}}_{\mathcal{L}}) &= g(|\mathcal{L}|)H(\mathbf{A}) \\ &\leq \alpha H(\mathbf{A}), \end{aligned}$$

where the proof for the equality is omitted due to space constraints; the inequality holds by (20) when $\alpha \geq \frac{\ell}{k}$, and in the case $\alpha < \frac{\ell}{k}$ it holds by (21), (22), and (23).

References

[1] A. C. Yao, "Protocols for secure computations," in *Proc. Annu. Symp. Found. Comput. Sci. (SFCS)*, pp. 160–164, IEEE, 1982.

[2] D. Chaum, I. B. Damgård, and J. Van de Graaf, "Multiparty computations ensuring privacy of each party's input and correctness of the result," in *Proc. Adv. Cryptol. (CRYPTO)*, pp. 87–119, Springer, 1988.

[3] M. Soleymani, H. Mahdavifar, and A. S. Avestimehr, "Analog secret sharing with applications to private distributed learning," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 1893–1904, 2022.

[4] S. Fu, Y. Yu, and M. Xu, "A secure algorithm for outsourcing matrix multiplication computation in the cloud," in *Proc. ACM Int. Workshop Secur. Cloud Comput. (SCC)*, pp. 27–33, ACM, 2017.

[5] J. Zhu and X. Tang, "Secure batch matrix multiplication from grouping Lagrange encoding," *IEEE Commun. Lett.*, vol. 25, no. 4, pp. 1119–1123, 2020.

[6] J. Kakar, S. Ebadifar, and A. Sezgin, "On the capacity and straggler-robustness of distributed secure matrix multiplication," *IEEE Access*, vol. 7, pp. 45783–45799, 2019.

[7] X. Jiang, M. Kim, K. Lauter, and Y. Song, "Secure outsourced matrix computation and application to neural networks," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, pp. 1209–1222, 2018.

[8] R. Gilad-Bachrach, N. Dowlin, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, "Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 201–210, PMLR, 2016.

[9] W.-T. Chang and R. Tandon, "On the capacity of secure distributed matrix multiplication," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, pp. 1–6, IEEE, 2018.

[10] Z. Jia and S. A. Jafar, "On the capacity of secure distributed batch matrix multiplication," *IEEE Trans. Inf. Theory*, vol. 67, no. 11, pp. 7420–7437, 2021.

[11] M. Aliasgari, O. Simeone, and J. Kliewer, "Private and secure distributed matrix multiplication with flexible communication load," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 2722–2734, 2020.

[12] J. Kakar, A. Khristoforov, S. Ebadifar, and A. Sezgin, "Uplink-downlink trade-off in secure distributed matrix multiplication," *arXiv preprint arXiv:1910.13849*, 2019.

[13] Z. Chen, Z. Jia, Z. Wang, and S. A. Jafar, "GCSA codes with noise alignment for secure coded multi-party batch matrix multiplication," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 1, pp. 306–316, 2021.

[14] H. H. López, G. L. Matthews, and D. Valvo, "Secure matdot codes: A secure, distributed matrix multiplication scheme," in *Proc. IEEE Inf. Theory Workshop (ITW)*, pp. 149–154, IEEE, 2022.

[15] R. G. D'Oliveira, S. El Rouayheb, and D. Karpuk, "GASP codes for secure distributed matrix multiplication," *IEEE Trans. Inf. Theory*, vol. 66, no. 7, pp. 4038–4050, 2020.

[16] Q. Yu and A. S. Avestimehr, "Coded computing for resilient, secure, and privacy-preserving distributed matrix multiplication," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 59–72, 2020.

[17] Q. Yu and A. S. Avestimehr, "Entangled polynomial codes for secure, private, and batch distributed matrix multiplication: Breaking the "cubic" barrier," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 245–250, IEEE, 2020.

[18] O. Makkonen, E. Saçıkara, and C. Hollanti, "Algebraic geometry codes for secure distributed matrix multiplication," *arXiv preprint arXiv:2303.15429*, 2023.

[19] R. A. Machado, G. L. Matthews, and W. Santos, "Hera scheme: Secure distributed matrix multiplication via hermitian codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 1729–1734, IEEE, 2023.

[20] A. B. Das, A. Ramamoorthy, and N. Vaswani, "Efficient and robust distributed matrix computations via convolutional coding," *IEEE Trans. Inf. Theory*, vol. 67, no. 9, pp. 6266–6282, 2021.

[21] B. Hasırcıoğlu, J. Gómez-Vilardebó, and D. Gündüz, "Bivariate polynomial coding for efficient distributed matrix multiplication," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 3, pp. 814–829, 2021.

[22] A. K. Pradhan, A. Heidarzadeh, and K. R. Narayanan, "Factored lt and factored raptor codes for large-scale distributed matrix multiplication," *IEEE J. Sel. Areas Inf. Theory*, vol. 2, no. 3, pp. 893–906, 2021.

[23] N. Mital, C. Ling, and D. Gündüz, "Secure distributed matrix computation with discrete Fourier transform," *IEEE Trans. Inf. Theory*, vol. 68, no. 7, pp. 4666–4680, 2022.

[24] H. A. Nodehi and M. A. Maddah-Ali, "Limited-sharing multi-party computation for massive matrix operations," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 1231–1235, IEEE, 2018.

[25] R. Bitar, M. Egger, A. Wachter-Zeh, and M. Xhemrishi, "Sparsity and privacy in secret sharing: A fundamental trade-off," *IEEE Trans. Inf. Forensics Secur.*, 2024.

[26] Z. Jia, H. Sun, and S. A. Jafar, "Cross subspace alignment and the asymptotic capacity of x-secure t-private information retrieval," *IEEE Trans. Inf. Theory*, vol. 65, no. 9, pp. 5783–5798, 2019.

[27] A. Beimel, "Secret-sharing schemes: A survey," in *Proc. Int. Conf. Coding Cryptol.*, pp. 11–46, Springer, 2011.

[28] A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, pp. 612–613, 1979.

[29] H. Yamamoto, "Secret sharing system using (k, l, n) threshold scheme," *Electronics Commun. Jpn. Part I Commun.*, vol. 69, no. 9, pp. 46–54, 1986.

[30] R. A. Chou and J. Kliewer, "Secure distributed storage: Optimal trade-off between storage rate and privacy leakage," *IEEE Trans. Inf. Theory*, 2024.

[31] M. Keshvari and R. A. Chou, "Distributed storage over a public channel: Trade-off between privacy and shared key lengths," in *Proc. Allerton Conf. Commun. Control Comput.*, pp. 1–5, 2023.

[32] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 1999.