SIAM J. APPL. ALGEBRA GEOMETRY
Vol. 8, No. 3, pp. 553–582

The Service Rate Region Polytope*

Gianira N. Alfarano†, Altan B. Kılıç†, Alberto Ravagnani†, and Emina Soljanin‡

Abstract. We investigate the properties of a family of polytopes that naturally arise in connection with a problem in distributed data storage, namely service rate region polytopes. The service rate region of a distributed coded system describes the data access requests that the underlying system can support. In this paper, we study the polytope structure of the service rate region with the primary goal of describing its geometric shape and properties. We achieve this by introducing various structural parameters of the service rate region and establishing upper and lower bounds for them. The techniques we apply in this paper range from coding theory to optimization. One of our main results shows that every rational point of the service rate region has a so-called rational allocation, answering an open question in the research area.

Key words. service rate region, polytope, error-correcting code, distributed data storage, linear programming, combinatorial geometry, convex polytopes, erasure-coded systems

MSC codes. 94B05, 52B12, 68P20, 52B55

DOI. 10.1137/23M1557829

Introduction. Distributed storage systems split data across servers to provide access services to multiple, possibly concurrent, users. The simplest way to reliably handle concurrent requests is to replicate data according to their popularity; see, for instance, [23, 28]. Unfortunately, this method can be expensive in terms of storage. Moreover, predicting how the interest in data changes is not always easy. For these reasons, erasure-coding has gained attention as a form of redundant storage; see, e.g., [10] and references therein.

Recent work establishes the concept of *service rate region* as an essential measure of the efficiency of a distributed coded system that should be considered in the system's design phase; see [1, 2, 4, 16, 17, 18]. To understand this metric, consider distributed systems in which k data objects are stored, using a linear $[n,k]_q$ error-correcting code across n servers, each with the same capacity $\mu \in \mathbb{R}$. The *service rate region* of the distributed coded system is the set of all request rates $(\lambda_1, \dots, \lambda_k) \in \mathbb{R}^k$ that the system can simultaneously handle. Such a distributed system is defined by a full rank $k \times n$ matrix G, and its service rate region is a convex polytope in $\mathbb{R}^k$.

*Received by the editors March 8, 2023; accepted for publication (in revised form) April 26, 2024; published electronically August 13, 2024.
https://doi.org/10.1137/23M1557829
Funding: The first author's work was supported by Swiss National Foundation grant 210966. The second author's work was supported by Dutch Research Council grant VI.Vidi.203.045. The third author's work was supported by Dutch Research Council grants VI.Vidi.203.045 and OCENW.KLEIN.539 and by the Royal Academy of Arts and Sciences of The Netherlands. The fourth author's work was supported in part by NSF award CIF-2122400.
†Eindhoven University of Technology, The Netherlands (gianira.alfarano@gmail.com, a.b.kilic@tue.nl, a.ravagnani@tue.nl).
‡Rutgers University, Piscataway, NJ 08854 USA (emina.soljanin@rutgers.edu).

Previous work in the area focuses on characterizing the service rate region of a distributed coded system and finding the optimal strategy to split the rate requests across the servers to maximize the region; see, e.g., [2]. The service rate region has been characterized for *binary simplex codes* and two classes of *maximum distance separable* (MDS) *codes*: (1) *systematic* MDS codes when $n \geq 2k$ and (2) MDS codes with arbitrary length and dimension that do not permit any data objects decoding from fewer than k stored objects; see again [2] and references therein.

A combinatorial approach to the service rate region has been introduced in [16], establishing and using the equivalence between the service rate problem and the well-known fractional matching problem on hypergraphs. In the same work, the authors showed that the service rate problem generalizes, in a precise sense, *batch, private information retrieval* (PIR), and *switch codes*; see [11, 12, 15, 25, 29, 30] for the details and background material about these classes of codes. In [17], the service rate regions of the *binary first-order Reed–Muller codes* and *binary simplex codes* have been determined using a geometric approach. In [3], coding-theoretic tools have been used to identify a polytope that contains the service rate region, giving an outer bound for it.

Contributions. In contrast with previous approaches, this paper focuses on describing the polytope structure of the service rate region and its geometric properties, such as its volume. Key tools to achieve this goal are outer bounds for the service rate region polytope, which we obtain by applying methods ranging from coding theory to convex geometry.

We also propose a discretized notion of service rate region that arises naturally in the integer allocation model. We prove that the discretized notion returns precisely the rational points of the originally proposed (continuous) notion. When investigating the connection between the allocation and the service rate region polytopes, we show that every rational point of the region has a rational allocation. The last part of the paper focuses on the service rate regions of systematic MDS matrices, for which we compute, for example, the volume in dimensions 2 and 3.

The rest of the paper is organized as follows: Section 1 introduces access models and states the service rate problem. Section 2 is devoted to the various representations of the service rate region polytope. In section 3, we discretize the concept of service rate region, also showing that rational points in the region have rational allocation. In section 4, we introduce and study the rth max-sum capacity and the system's volume. Section 5 is devoted to proving different outer bounds on the service rate region. Section 6 focuses on systematic MDS codes. Elementary background on polytopes and error-correcting codes is provided in the appendices.

1. System model and problem formulation.

We consider a distributed service system with n identical nodes (servers). Each node has two functional components: one for data storage and the other for processing download requests posed by the system users. This section establishes the notation, defines distributed service systems and their service rate regions, and states the problems this paper addresses.

Throughout the paper, q denotes a prime power, and $\mathbb{F}_q$ is the finite field with q elements. We work with integers $n > k \geq 2$. All vectors in what follows are row vectors unless otherwise stated.

1.1. Storage model. We consider a coded, distributed data storage system where k objects (elements of $\mathbb{F}_q$) are linearly encoded and stored across n servers. Each server stores precisely one element of $\mathbb{F}_q$. Therefore, the coded system is specified by a rank k matrix $G \in \mathbb{F}_q^{k\times n}$, which we call the *generator matrix* of the system. If $(x_1,\dots,x_k) \in \mathbb{F}_q^k$ is the k-tuple of objects to be stored, then the jth server stores the jth component of the encoded vector

$$(x_1,\dots,x_k)\cdot G \in \mathbb{F}_q^n.$$

Note that, by definition, the n servers store *linear combinations* of the data objects rather than just copies of them. The latter storage strategy is called (*simple*) *replication.*

Following the coding theory terminology [21], we say that the matrix G is *systematic* if its first k columns form the identity $k\times k$ matrix. If the νth column of G is a nonzero multiple of the standard basis vector e_i, then we say that ν is a *systematic node* for the ith data object. If every $\nu \in \{1,\dots,n\}$ is a systematic node, then we say that G is a *replication matrix.* Note that a replication matrix describes a system where each object is stored as it is (up to nonzero multiples).

Notation 1.1. *To simplify the statements throughout the paper, without loss of generality, we work with a fixed matrix $G \in \mathbb{F}_q^{k\times n}$ of rank k. We denote by G^ν the νth column of G and assume that none of its columns is the zero vector.*

Since $n > k$, we may be able to recover each object from different sets of servers, which motivates the following definitions and terminology.

Definition 1.2. *Let $R \subseteq \{1,\dots,n\}$ be such that $e_i \in \langle G^\nu \mid \nu \in R\rangle$ and $\langle G^\nu \mid \nu \in R\rangle$ is the $\mathbb{F}_q$-span of the columns of G indexed by R. Then R is called a* recovery set *for the ith object. For $i \in \{1,\dots,k\}$, let*

$$\mathcal{R}_i^{\text{all}}(G) := \{R \subseteq \{1,\dots,n\} \mid e_i \in \langle G^\nu \mid \nu \in R\rangle\},$$

where superscript "all" indicates that $\mathcal{R}_i^{\text{all}}(G)$ contains all the recovery sets for the ith object.

Since G has rank k by assumption, we have $\mathcal{R}_i^{\text{all}}(G) \neq \emptyset$ for all $i \in \{1,\dots,k\}$. Moreover, $R \neq \emptyset$ for all $i \in \{1,\dots,k\}$ and $R \in \mathcal{R}_i^{\text{all}}(G)$. We continue by formalizing the concept of a recovery system.

Definition 1.3. *A (recovery) G-system is a k-tuple $\mathcal{R} = (\mathcal{R}_1,\dots,\mathcal{R}_k)$, with $\mathcal{R}_i \subseteq \mathcal{R}_i^{\text{all}}(G)$ and $\mathcal{R}_i \neq \emptyset$ for all $i \in \{1,\dots,k\}$.*

1.2. Service and access models. We adopt two service models and refer to them as the *queuing* and the *bandwidth* models. They are based on two common ways of implementing resource sharing by incoming data access requests. In the queueing model, the requests to download from a node are placed in its queue. Each node can serve on average μ requests per unit of time. To maintain the stability of each queue, the total request arrival rate at each node must not exceed its service rate μ.

In the bandwidth model, each node can concurrently serve multiple data access requests. When each node has an I/O bus with a finite access bandwidth W bits/second and a download requires streaming at a fixed bandwidth of b bits/second, a node can simultaneously serve up to $\mu = \lfloor W/b \rfloor$ number of requests. In both cases, we refer to μ as the *server's capacity*

(formal definitions will be given later). In the queuing model, requests to download object i arrive at rate $\lambda_i \in \mathbb{R}_{\geq 0}$. In the bandwidth model, λ_i is the number of object i requests simultaneously in the system. In both models, $\lambda_{i,R}$ is the portion of λ_i assigned to be served by the recovery set $R \in \mathcal{R}_i$. We refer to a set $\{\lambda_{i,R} \mid R \in \mathcal{R}_i,\, i = 1, \ldots, k\}$ as a *request allocation*.

1.3. Normalization and integrality. We can normalize all request allocation values and rates by dividing them by the node service capacity μ. In this case, all normalized request rates λ_i and the numbers in $\{\lambda_{i,R} \mid R \in \mathcal{R}_i,\, i \in \{1, \ldots, k\}\}$ are multiples of $1/\mu$. Since in the bandwidth model, these numbers count requests, their normalized versions are integer multiples of $1/\mu$. Furthermore, there are practical scenarios wherein each served request occupies the entire bandwidth of the server it is accessing (e.g., streaming from low-bandwidth edge devices). In such cases, λ_i are integers, and $\lambda_{i,R}$ are binary numbers. In other practical scenarios, a user can simultaneously download data from multiple nodes at a fraction of its bandwidth from each, and the assumption that $\lambda_{i,R}$ is integer multiples of $1/\mu$ can be relaxed.

1.4. Service rate region and problem formulation. We are interested in characterizing the k-tuples $(\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^k$ of rate requests that the data storage system can support. The set of such tuples is formally defined as follows, yielding to the notion of the *service rate region* of a distributed storage system.

Definition 1.4. *Let $\mathcal{R} = (\mathcal{R}_1, \ldots, \mathcal{R}_k)$ be a G-system. The* service rate region *associated with $\mathcal{R}$ and μ is the set of all $(\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^k$ for which there exists a collection of real numbers*

$$\{\lambda_{i,R} \mid i \in \{1, \ldots, k\},\, R \in \mathcal{R}_i\}$$

with the following properties:

$$\sum_{R \in \mathcal{R}_i} \lambda_{i,R} = \lambda_i \text{ for } 1 \leq i \leq k, \tag{1.1}$$

$$\sum_{i=1}^{k} \sum_{\substack{R \in \mathcal{R}_i \\ \nu \in R}} \lambda_{i,R} \leq \mu \text{ for } 1 \leq \nu \leq n, \tag{1.2}$$

$$\lambda_{i,R} \geq 0 \text{ for } 1 \leq i \leq k,\, R \in \mathcal{R}_i. \tag{1.3}$$

A collection $\{\lambda_{i,R}\}$ that satisfies properties (1.2) and (1.3) above is called a feasible allocation *for the pair $(\mathcal{R}, \mu)$. The service rate region associated with $\mathcal{R}$ and μ is denoted by*

$$\Lambda(\mathcal{R}, \mu) \subseteq \mathbb{R}^k.$$

Observe that the service rate region of a G-system is independent of the ordering of the recovery sets in each $\mathcal{R}_i$.

Remark 1.5. It turns out that $\Lambda(\mathcal{R}, \mu)$ is a down-monotone polytope; see Theorem 2.3 below and Appendix A for the definitions. We will elaborate on this when introducing the allocation polytope in section 2.

The service rate region of a G-system $\mathcal{R}$ may not change if we select a suitable subset of the recovery sets, which allows us to reduce the number of variables and inequalities in Definition 1.4. We start with the following observation, whose proof is simple and therefore omitted.

Proposition 1.6. *Suppose that $\mathcal{R}=(\mathcal{R}_1,\dots,\mathcal{R}_k)$ and $\mathcal{R}'=(\mathcal{R}'_1,\dots,\mathcal{R}'_k)$ are G-systems with $\mathcal{R}'_i\subseteq\mathcal{R}_i$ for all $i\in\{1,\dots,k\}$. Then $\Lambda(\mathcal{R},\mu)\supseteq\Lambda(\mathcal{R}',\mu)$. In particular, $\Lambda(\mathcal{R},\mu)\subseteq\Lambda(\mathcal{R}^{\mathrm{all}}(G),\mu)$ for any G-system $\mathcal{R}$.*

The service rate region $\Lambda(\mathcal{R}^{\mathrm{all}}(G),\mu)$ does not change when we select from $\mathcal{R}^{\mathrm{all}}(G)$ the recovery sets that are minimal with respect to inclusion, in the following precise sense.

Definition 1.7. *A set $R\in\mathcal{R}_i^{\mathrm{all}}$ is called* i-minimal *if there is no $R'\in\mathcal{R}_i^{\mathrm{all}}(G)$ with $R'\subsetneq R$. We let $\mathcal{R}^{\mathrm{min}}(G)$ be the G-system defined, for all i, by*

$$\mathcal{R}_i^{\mathrm{min}}(G):=\{R\in\mathcal{R}_i^{\mathrm{all}}(G)\mid R\text{ is }i\text{-minimal}\}.$$

The proof of the following result is not difficult and is therefore left to the reader.

Proposition 1.8. *We have $\Lambda(\mathcal{R}^{\mathrm{min}}(G),\mu)=\Lambda(\mathcal{R}^{\mathrm{all}}(G),\mu)$.*

Remark 1.9. It immediately follows from the definitions that $\Lambda(\mathcal{R},\mu)=\mu\Lambda(\mathcal{R},1)$ for any G-system $\mathcal{R}$, where $\mu\Lambda(\mathcal{R},1)=\{\mu\lambda\mid\lambda\in\Lambda(\mathcal{R},1)\}$. In what follows, we will often assume $\mu=1$ without loss of generality.

The following symbols will further simplify the statements in what follows.

Notation 1.10. *For a G-system $\mathcal{R}=(\mathcal{R}_1,\dots,\mathcal{R}_k)$ we let $\Lambda(\mathcal{R})=\Lambda(\mathcal{R},1)$. We also write $\Lambda(G,\mu)=\Lambda(\mathcal{R}^{\mathrm{all}}(G),\mu)=\Lambda(\mathcal{R}^{\mathrm{min}}(G),\mu)$, where the latter equality follows from Proposition 1.8. Finally, we set $\Lambda(G)=\Lambda(G,1)$.*

We conclude with an example illustrating the concepts introduced in this section.

Example 1.11. Consider the matrices

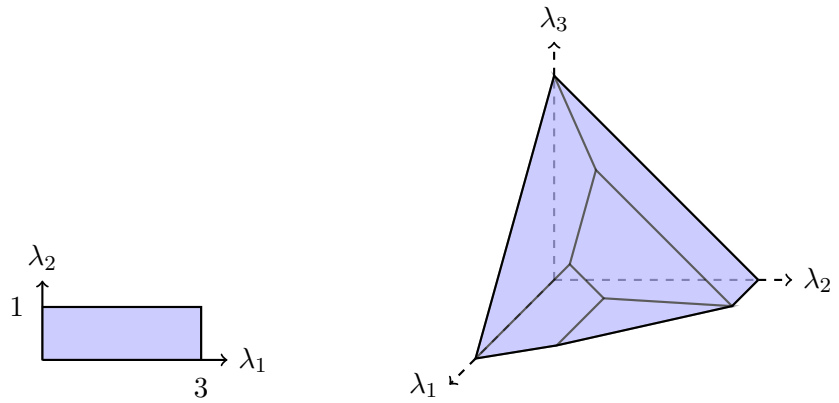
$$G_1=\begin{pmatrix}1&0&1&1\\0&1&0&0\end{pmatrix}\in\mathbb{F}_2^{2\times4},\qquad G_2=\begin{pmatrix}1&0&0&1&0&1\\0&1&0&1&2&2\\0&0&1&1&1&1\end{pmatrix}\in\mathbb{F}_3^{3\times6}.$$

The corresponding service rate regions are depicted in Figure 1. For the matrix G_2 we have

$$\begin{aligned}\mathcal{R}_1^{\mathrm{min}}(G_2)&=\{\{1\},\{5,6\},\{2,3,4\},\{2,4,5\},\{3,4,6\},\{2,3,6\},\{3,4,5\}\},\\ \mathcal{R}_2^{\mathrm{min}}(G_2)&=\{\{2\},\{3,5\},\{4,6\},\{1,3,4\},\{1,4,5\},\{1,3,6\}\},\\ \mathcal{R}_3^{\mathrm{min}}(G_2)&=\{\{3\},\{2,5\},\{1,2,4\},\{1,4,6\},\{1,2,6\},\{1,4,5\}\}.\end{aligned}$$

Moreover, for all $i\in\{1,2,3\}$ we have $\mathcal{R}_i^{\mathrm{all}}=\{R\subseteq\{1,\dots,6\}\mid S\subseteq R\text{ for some }S\in\mathcal{R}_i^{\mathrm{min}}(G_2)\}$. Finally, to see that, for example, the point $P=(3/2,3/2,1/2)$ belongs to $\Lambda(G_2)$, we can consider the feasible allocation given by

$$\lambda_{1,R}=\begin{cases}1&\text{if }R=\{1\},\\1/2&\text{if }R=\{5,6\},\\0&\text{otherwise,}\end{cases}\qquad\lambda_{2,R}=\begin{cases}1&\text{if }R=\{2\},\\1/2&\text{if }R=\{3,5\},\\0&\text{otherwise,}\end{cases}\qquad\lambda_{3,R}=\begin{cases}1/2&\text{if }R=\{3\},\\0&\text{otherwise.}\end{cases}$$



(a) $\Lambda(G_1)$, Example 1.11. (b) $\Lambda(G_2)$, Example 1.11.

Figure 1. *The service rate regions of the systems in Example 1.11.*

It is easy to see that the collection $\{\lambda_{i,R}\}$ satisfies the properties (1.1)–(1.3) for $(\mathcal{R}^{\min}(G_2), 1)$.

This paper mainly describes the geometric properties of the service rate region $\Lambda(\mathcal{R}, \mu)$. Most of our results hold for an arbitrary G-system $\mathcal{R}$, although our main focus is on $\Lambda(G)$.

2. The service rate region and the allocation polytopes. This section describes the polytope structure of the service rate region associated with an arbitrary G-system $\mathcal{R}$. We also illustrate how the geometric structure has implications for the allocation of users in the corresponding system. For these purposes, viewing the service rate region as the image of a higher dimensional polytope under a linear map is often convenient; we call this the *allocation polytope.*

Definition 2.1. *Let $\mathcal{R} = (\mathcal{R}_1, \dots, \mathcal{R}_k)$ be a G-system, $m_i = |\mathcal{R}_i|$ for $i \in \{1, \dots, k\}$, and $m(\mathcal{R}) = m_1 + \cdots + m_k$. The* allocation polytope *of $(\mathcal{R}, \mu)$ is the set of $(\lambda_{i,R} \mid i \in \{1, \dots, k\}, R \in \mathcal{R}_i)$ that satisfy the inequalities (1.2) and (1.3). We denote the allocation polytope by*

$$\mathcal{A}(\mathcal{R}, \mu) \subseteq \mathbb{R}^{m(\mathcal{R})}.$$

We also let $\mathcal{A}(G) = \mathcal{A}(\mathcal{R}^{\text{all}}(G)) = \mathcal{A}(\mathcal{R}^{\min}(G))$, where the latter identity can be shown similarly to Proposition 1.8.

We now show that the allocation polytope is indeed a polytope, and we state its connection with the service rate region. We will use the following fact, which easily follows from the definition of a convex hull combined with Theorem A.1 and the observations right after it.

Lemma 2.2. *Let $\mathcal{P} \subseteq \mathbb{R}^m$ be a polytope and let $f : \mathbb{R}^m \to \mathbb{R}^k$ be a linear map. Then $f(\mathcal{P})$ is a polytope and $|(f(\mathcal{P})) \subseteq f(|(\mathcal{P}))$.*

The connection between the allocation polytope and the service rate region is described by the next result, which also summarizes some properties of these two regions. In particular, we are interested in maps of the form given in Theorem 2.3.

Theorem 2.3. *Let $\mathcal{R}$ be a G-system and let $m(\mathcal{R})$ be as in Definition 2.1. The following hold.*

1. *We have* $f(\mathcal{A}(\mathcal{R},\mu)) = \Lambda(\mathcal{R},\mu)$*, where* $f : \mathbb{R}^{m(\mathcal{R})} \to \mathbb{R}^k$ *is the linear map defined by*

$$f : \lambda = (\lambda_{i,R} | i \in \{1,\dots,k\},\, R \in \mathcal{R}_i) \mapsto \left(\sum_{R\in\mathcal{R}_1} \lambda_{1,R}, \dots, \sum_{R\in\mathcal{R}_k} \lambda_{k,R} \right).$$

2. $\mathcal{A}(\mathcal{R},\mu)$ *and* $\Lambda(\mathcal{R},\mu)$ *are down-monotone polytopes.*

Proof. The fact that $\Lambda(\mathcal{R},\mu)$ is the image of $\mathcal{A}(\mathcal{R},\mu)$ under f easily follows from Definitions 1.4 and 2.1. We now establish the second part of the statement. The set $\mathcal{A}(\mathcal{R},\mu)$ is a polyhedron by definition. Its boundedness can be shown as follows. Summing all the inequalities in (1.2), we get $\sum_{i=1}^k \sum_{R\in\mathcal{R}_i} \lambda_{i,R} \le n\mu$. Using $\lambda_{i,R} \ge 0$ for all pairs (i,R), we get that any $\lambda \in \mathcal{A}(\mathcal{R},\mu)$ satisfies

$$\|\lambda\|_2 \le \sum_{i=1}^k \sum_{R\in\mathcal{R}_i} |\lambda_{i,R}| = \sum_{i=1}^k \sum_{R\in\mathcal{R}_i} \lambda_{i,R} \le n\mu,$$

where $\|\lambda\|_2$ is the 2-norm of λ. The fact that $\Lambda(\mathcal{R},\mu)$ is bounded follows from Lemma 2.2 and the boundedness of $\mathcal{A}(\mathcal{R},\mu)$. Finally, it is not hard to directly check that the polytopes $\mathcal{A}(\mathcal{R},\mu)$ and $\Lambda(\mathcal{R},\mu)$ are down-monotone. ∎

We now turn to the natural question of describing the vertices and the points of the service rate region with rational entries. We will use the connection between the service rate region and the allocation polytope to answer these questions.

We note that a linear map $f : \mathbb{R}^m \to \mathbb{R}^k$ is called *rational* if its matrix concerning the canonical basis has rational entries. The next result follows from Proposition A.2 and Corollary A.3.

Lemma 2.4. *Let* $\mathcal{P} \subseteq \mathbb{R}^m$ *be a rational polytope and* $f : \mathbb{R}^m \to \mathbb{R}^k$ *be a rational linear map. Then* $f(\mathcal{P})$ *is a polytope whose vertices have rational entries.*

The following result will be crucial to qualitatively describe the connection between the allocation polytope and the service rate region.

Lemma 2.5. *Let* $\mathcal{P} \subseteq \mathbb{R}^m$ *be a rational polytope, and let* $f : \mathbb{R}^m \to \mathbb{R}^k$ *be a rational linear map. We have* $f(\mathcal{P} \cap \mathbb{Q}^m) = f(\mathcal{P}) \cap \mathbb{Q}^k$.

Proof. The inclusion $\subseteq$ is immediate. To prove the other inclusion, we let $y \in f(\mathcal{P}) \cap \mathbb{Q}^k$. Write $f = (f_1,\dots,f_k)$, where $f_i : \mathbb{R}^m \to \mathbb{R}$. We need to show that there exists $x \in \mathcal{P} \cap \mathbb{Q}^m$ with $f_i(x) = y_i$ for all $i \in \{1,\dots,k\}$. Since $\mathcal{P}$ is rational, $\mathcal{P} = \{x \in \mathbb{R}^m \mid Ax^\top \le b^\top\}$ for some $A \in \mathbb{Q}^{\ell\times m}$ and $b \in \mathbb{Q}^\ell$. We append to A and b a total of $2k$ rows, of which k are for the inequalities $\{f_i(x) \le y_i \mid i = 1,\dots,k\}$ and the other k are for the inequalities $\{-f_i(x) \le -y_i \mid i = 1,\dots,k\}$. Let A' and b' denote the resulting matrix and vector, of size $(\ell+2k)\times m$ and length $\ell+2k$, respectively. Then, by definition, $\mathcal{P}' = \{x \in \mathbb{R}^m \mid A'x^\top \le b'^\top\}$ is a polyhedron. Note that we need to prove that $\mathcal{P}' \cap \mathbb{Q}^m \ne \emptyset$. We first observe that $\mathcal{P}'$ is bounded because $\mathcal{P}' \subseteq \mathcal{P}$ and $\mathcal{P}$ is bounded. Moreover, $\mathcal{P}'$ is rational because A' and b' have rational entries (here, we use the fact that y has rational entries and f is a rational map). Moreover, $\mathcal{P}'$ is nonempty because $y \in f(\mathcal{P})$ by assumption, which implies the existence of $z \in \mathcal{P}$ with $f(z) = y$, i.e., of $z \in \mathcal{P}'$. Then $\mathcal{P}'$ has at least one vertex x and that vertex must have rational entries by Corollary A.3. Thus $\mathcal{P}' \cap \mathbb{Q}^m \ne \emptyset$, as desired. ∎

We are now ready to state the main result of this section (note that the properties of the statement only hold for $\mu = 1$).

Theorem 2.6. *Let $\mathcal{R}$ be a G-system. The following hold.*

1. *The vertices of $\Lambda(\mathcal{R})$ have rational entries.*
2. *$\Lambda(\mathcal{R}) \cap \mathbb{Q}^k = f(\mathcal{A}(\mathcal{R}) \cap \mathbb{Q}^k)$, where f is defined as in Theorem 2.3.*

Proof. The vertices of $\Lambda(\mathcal{R})$ have rational entries because of Theorem 2.3 and Lemma 2.4. The second part of the statement follows by combining Theorem 2.3 with Lemma 2.5. ∎

Note that property 2 of Theorem 2.6 implies that every rational point of the service rate region has a feasible rational allocation; see Definition 1.4. This fact, which does not appear to be obvious, is important from the application point of view in the sense of subsection 1.3.

3. The integer allocation model.

In this section, we consider practical scenarios, described in section 1.3, wherein each server has a specific bandwidth, and each served request occupies the entire bandwidth when served, i.e., the $\lambda_{i,R}$'s are constrained to be either 0 or 1. We define the service rate region for this model and show how it relates to the model we considered in the previous section.

Assume we use the system $s \in \mathbb{Z}_{\geq 1}$ times, each time with possibly different allocation. Let $\alpha_i(R)$ be the number of times that recovery set $R \in \mathcal{R}_i$ is used to recover the ith object within the s uses of the system. Then the number of times the ith object is recovered is equal to $\lambda_i = \sum_{R \in \mathcal{R}_i} \alpha_i(R)$. This motivates the following definitions.

Definition 3.1. *Let $\mathcal{R}$ be a G-system. An $\mathcal{R}$-allocation is a k-tuple of functions $\alpha = (\alpha_1, \dots, \alpha_k)$, where $\alpha_i : \mathcal{R}_i \to \mathbb{N}$ for all $i \in \{1, \dots, k\}$. The service rate of α is the vector $\lambda(\alpha) = (\lambda_1, \dots, \lambda_k) \in \mathbb{N}^k$, where*

$$\lambda_i = \sum_{R \in \mathcal{R}_i} \alpha_i(R) \qquad \text{for all } i \in \{1, \dots, k\}.$$

For $\nu \in \{1, \dots, n\}$ we define

$$\delta_\nu(\mathcal{R}, \alpha) = \sum_{i=1}^{k} \sum_{R \in \mathcal{R}_i} \delta_\nu(R)\, \alpha_i(R) \text{ where } \delta_\nu(R) = \begin{cases} 1 & \text{if } \nu \in R, \\ 0 & \text{otherwise.} \end{cases}$$

In Definition 3.1, the quantity $\delta_\nu(\mathcal{R}, \alpha)$ represents the number of times server ν is contacted.

Definition 3.2. *Let $\mathcal{R}$ be a G-system. The* one-shot service rate region *of $\mathcal{R}$ with* capacity *$s \in \mathbb{Z}_{\geq 1}$ is $\Lambda_1(\mathcal{R}, s) = \{\lambda(\alpha)/s \mid \alpha$ an $\mathcal{R}$-allocation, $\delta_\nu(\mathcal{R}, \alpha) \leq s$ for $1 \leq \nu \leq n\}$. The* rational service rate region *of $\mathcal{R}$ is the set*

$$\Lambda^{\mathbb{Q}}(\mathcal{R}) = \bigcup_{s \in \mathbb{Z}_{\geq 1}} \Lambda_1(\mathcal{R}, s).$$

In this section, we will show the following "topological" connection between the rational service rate region and the service rate region as defined in section 1.

Theorem 3.3. *Let $\mathcal{R}$ be a G-system. The following hold.*
1. $\Lambda^{\mathbb{Q}}(\mathcal{R}) = \Lambda(\mathcal{R}) \cap \mathbb{Q}^k$.
2. $\Lambda(\mathcal{R}) = \overline{\Lambda^{\mathbb{Q}}(\mathcal{R})}$, *where the latter is the closure of $\Lambda^{\mathbb{Q}}(\mathcal{R})$ with respect to the Euclidean topology in $\mathbb{R}^k$.*

Remark 3.4. Before proving Theorem 3.3, we stress that it is particularly relevant for the practical scenarios described in section 1.3, which may require that allocations be integer or rational. It shows that (1) the rational points in the service rate region can be achieved with rational allocations, and (2) all points can be achieved by averaging over multiple system uses.

We will use the following result in the proof of Theorem 3.3.

Lemma 3.5. *Let $\mathcal{P} \subseteq \mathbb{R}^m$ be a down-monotone polytope. Then $\mathcal{P} = \overline{\mathcal{P} \cap \mathbb{Q}^m}$, where the latter is the closure of $\mathcal{P} \cap \mathbb{Q}^m$ for the Euclidean topology.*

Proof. The inclusion $\overline{\mathcal{P} \cap \mathbb{Q}^m} \subseteq \mathcal{P}$ holds because $\mathcal{P}$ is closed and $\overline{\mathbb{Q}^m} = \mathbb{R}^m$, which implies $\overline{\mathcal{P} \cap \mathbb{Q}^m} \subseteq \overline{\mathcal{P}} \cap \overline{\mathbb{Q}^m} = \mathcal{P} \cap \mathbb{R}^m = \mathcal{P}$. For the other inclusion, we will prove that for all $x \in \mathcal{P}$ and all $\varepsilon > 0$ we have $B_\varepsilon(x) \cap \mathcal{P} \cap \mathbb{Q}^m \neq \emptyset$, where $B_\varepsilon(x)$ is the ball of radius ε centered at x. This implies $\mathcal{P} \subseteq \overline{\mathcal{P} \cap \mathbb{Q}^m}$ using, for example, [22, Theorems 17.5 and 20.3]. Fix any x and ε as above. Write $x = (x_1, \dots, x_m) \in \mathbb{R}^m$. Since $\mathbb{Q}$ is dense in $\mathbb{R}$, for every $i \in \{1, \dots, m\}$ there exists $y_i \in \mathbb{Q}$ with $x_i - \varepsilon/m \leq y_i \leq x_i$. Since $\mathcal{P}$ is down-monotone, we have $y = (y_1, \dots, y_m) \in \mathcal{P}$. The fact that $y \in B_\varepsilon(x) \cap \mathcal{P} \cap \mathbb{Q}^m$ now follows from

$$\|x - y\|_2 \leq \|x - y\|_1 = \sum_{i=1}^{m} (x_i - y_i) \leq m \cdot \varepsilon/m = \varepsilon,$$

We used the standard notation for the p-norm in $\mathbb{R}^k$. ∎

Proof of Theorem 3.3. The second part of the statement follows from the first part in combination with Lemma 3.5. Therefore it suffices to establish the first part.

Let $\lambda \in \Lambda^{\mathbb{Q}}(\mathcal{R})$. There is an $s \in \mathbb{Z}_{\geq 1}$ and an $\mathcal{R}$-allocation α such that $s\lambda = \lambda(\alpha) = (\lambda_1, \dots, \lambda_k) \in s\Lambda_1(\mathcal{R}, s)$ and $\delta_\nu(\mathcal{R}, \alpha) \leq s$ for all $\nu \in \{1, \dots, n\}$. Note that $\lambda \in \mathbb{Q}^k$. We will show that the set

$$\left\{ \frac{\alpha_i(R)}{s} \mid i \in \{1, \dots, n\},\ R \in \mathcal{R}_i \right\}$$

satisfies properties (1.1), (1.2), and (1.3). By definition, for $1 \leq i \leq k$ we have $\lambda_i/s = \sum_{R \in \mathcal{R}_i} \alpha_i(R)/s$. Moreover, the condition $\delta_\nu(\mathcal{R}, \alpha) \leq s$ can be rewritten as

$$\sum_{i=1}^{k} \sum_{\substack{R \in \mathcal{R}_i \\ \nu \in R}} \frac{\alpha_i(R)}{s} \leq 1 \quad \text{for } 1 \leq \nu \leq n.$$

Finally, $\alpha_i(R)/s \geq 0$ for all $i \in \{1, \dots, k\}$ and all $R \in \mathcal{R}_i$. We therefore conclude that $\lambda \in \Lambda(\mathcal{R}) \cap \mathbb{Q}^k$; hence $\Lambda^{\mathbb{Q}}(\mathcal{R}) \subseteq \Lambda(\mathcal{R}) \cap \mathbb{Q}^k$.

To prove the other containment, let $\lambda \in \Lambda(\mathcal{R}) \cap \mathbb{Q}^k$. By Theorem 2.6, we have $\Lambda(\mathcal{R}) \cap \mathbb{Q}^k = f(\mathcal{A}(\mathcal{R}) \cap \mathbb{Q}^k)$, where f is defined as in Theorem 2.3. In particular, there exist rational numbers $\{\lambda_{i,R} \in \mathbb{Q} \mid i \in \{1, \dots, k\},\ R \in \mathcal{R}_i\}$ that satisfy properties (1.1)–(1.3) of Definition 1.4. By definition, $\lambda_{i,R} = u_{i,R}/v_{i,R}$, $u_{i,R}, v_{i,R} \in \mathbb{N}$, and $v_{i,R} > 0$ for all $i \in \{1, \dots, k\}$ and $R \in \mathcal{R}_i$.

Let $s := \operatorname{lcm}(v_{i,R} \mid i \in \{1,\dots,k\},\ R \in \mathcal{R}_i)$, so that $s\lambda = (s\lambda_1,\dots,s\lambda_k) \in \mathbb{N}^k$. We claim that $s\lambda \in s\Lambda_1(\mathcal{R},s)$. Now for $1 \le i \le k$, define $\alpha_i(R) = s\lambda_{i,R}$; note that α_i always maps to $\mathbb{N}$ by the construction of the number s. Then for $1 \le i \le k$ we have $s\lambda_i = \sum_{R\in\mathcal{R}_i} \alpha_i(R)$, from which we conclude that $\alpha = (\alpha_1,\dots,\alpha_k)$ is an $\mathcal{R}$-allocation. Furthermore, for $\nu \in \{1,\dots,n\}$ we have

$$\delta_\nu(\mathcal{R},\alpha) = \sum_{i=1}^{k} \sum_{R\in\mathcal{R}_i} \delta_\nu(R)\alpha_i(R) = \sum_{i=1}^{k} \sum_{\substack{R\in\mathcal{R}_i \\ \nu\in R}} \alpha_i(R) = \sum_{i=1}^{k} \sum_{\substack{R\in\mathcal{R}_i \\ \nu\in R}} s\lambda_{i,R} \le s,$$

where the second equality follows from the definition of $\delta_\nu(R)$ and the last inequality follows from (1.2). This shows that $s\lambda \in s\Lambda_1(\mathcal{R},s)$, and equivalently $\lambda \in \Lambda_1(\mathcal{R},s)$; hence $\lambda \in \Lambda^{\mathbb{Q}}(\mathcal{R})$, as desired. ∎

We conclude this section with an example illustrating Theorem 3.3.

Example 3.6. Let

$$G := \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix} \in \mathbb{F}_3^{2\times 4}.$$

Consider the G-system $\mathcal{R} = \mathcal{R}^{\min}(G) = (\mathcal{R}_1,\mathcal{R}_2)$, where $\mathcal{R}_1 := \{\{1\},\{2,3\},\{2,4\},\{3,4\}\}$ and $\mathcal{R}_2 := \{\{2\},\{1,3\},\{1,4\},\{3,4\}\}$. The corresponding service rate region is depicted in Figure 2, along with the point $P = (4/3, 2/3)$.

We have $P \in \Lambda(\mathcal{R}) \cap \mathbb{Q}^2$ and $P \in \Lambda_1(\mathcal{R},3)$; i.e., P can be achieved in three uses of the system. An example of an $\mathcal{R}$-allocation, in the sense of Definition 3.1, is given by

$$\begin{aligned} \alpha_1 : \mathcal{R}_1 &\to \mathbb{N} & \qquad \alpha_2 : \mathcal{R}_2 &\to \mathbb{N} \\ \{1\} &\mapsto 2 & \{2\} &\mapsto 1 \\ \{2,3\} &\mapsto 1 & \{1,3\} &\mapsto 0 \\ \{2,4\} &\mapsto 0 & \{1,4\} &\mapsto 1 \\ \{3,4\} &\mapsto 1 & \{3,4\} &\mapsto 0 \end{aligned}$$

We have $\delta_1(\mathcal{R},\alpha) = \alpha_1(\{1\}) + \alpha_2(\{1,4\}) = 3$, $\delta_2(\mathcal{R},\alpha) = \alpha_1(\{2,3\}) + \alpha_2(\{2\}) = 2$, $\delta_3(\mathcal{R},\alpha) = \alpha_1(\{2,3\}) + \alpha_1(\{3,4\}) = 2$, $\delta_4(\mathcal{R},\alpha) = \alpha_1(\{3,4\}) + \alpha_2(\{1,4\}) = 2$. Moreover, $\sum_{R\in\mathcal{R}_1} \alpha_1(R) = 4$ and $\sum_{R\in\mathcal{R}_2} \alpha_2(R) = 2$, showing that $(4,2) \in \Lambda_1(\mathcal{R},3)$.

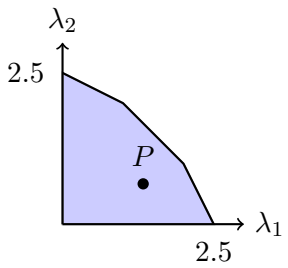


Figure 2. *Service rate region for the G-system in Example 3.6 and the point $P = (4/3, 2/3)$.*

4. Fundamental parameters of the service rate region. This section introduces some fundamental parameters of the service rate region that describe its "shape." We then study them by applying various techniques. The results are aimed at describing how the *algebra* of the underlying matrix G determines the *geometry* of the service region polytope $\Lambda(G,\mu)$. To simplify the notation (and without loss of generality), we assume $\mu = 1$. We start by recalling two types of elementary polytopes.

Definition 4.1. *Let $h, \delta \in \mathbb{R}_{\geq 0}$. The* h-hypercube *in $\mathbb{R}^k$ is the convex hull of the set $\{x \in \mathbb{R}^k_{\geq 0} \mid x_i \in \{0, h\}$ for $1 \leq i \leq k\}$. We say that h is the* size *of the hypercube. The* δ-simplex *in $\mathbb{R}^k$ is the convex hull of the set $\{\delta e_1, \ldots, \delta e_k\}$, where e_i is the ith standard basis vector of $\mathbb{F}_q^k$. Again, we say that δ is the* size *of the simplex.*

We will introduce the first set of parameters for the service rate region. Other parameters will be introduced later.

Definition 4.2. *Let $\mathcal{R}$ be a G-system. We let:*

$$\lambda^r(\mathcal{R}) = \max\left\{\sum_{i=1}^{k} \lambda_i^r \mid \lambda \in \Lambda(\mathcal{R})\right\}, \qquad [r\text{th max-sum capacity}]$$

$$\lambda(\mathcal{R}) = \lambda^1(\mathcal{R}), \qquad [\text{max-sum capacity}]$$

$$\lambda_i^*(\mathcal{R}) = \max\{x \in \mathbb{R} \mid xe_i \in \Lambda(\mathcal{R})\} \text{ for } 1 \leq i \leq k,$$

$$\lambda^*(\mathcal{R}) = \max\{\lambda_i^*(\mathcal{R}) \mid 1 \leq i \leq k\}.$$

Furthermore, we denote as follows the largest size of a hypercube and a simplex contained in the service rate region:

$$h(\mathcal{R}) = \max\{x \in \mathbb{R} \mid (x, \ldots, x) \in \Lambda(\mathcal{R})\},$$

$$\delta(\mathcal{R}) = \min\{\lambda_i^*(\mathcal{R}) \mid 1 \leq i \leq k\}.$$

When $\mathcal{R} = \mathcal{R}^{\mathrm{all}}(G)$ or $\mathcal{R} = \mathcal{R}^{\min}(G)$, we simply write $\lambda^r(G)$, $\lambda(G)$, $\lambda_i^*(G)$, $\lambda^*(G)$, $h(G)$, and $\delta(G)$.

The next example shows that, in general, a point achieving the max-sum capacity will not achieve the rth max-sum capacity for $r > 1$.

Example 4.3. Consider the service rate region of Example 3.6, depicted in Figure 2. One can show that $\lambda(G)$ is achieved by $(1,2)$ and $(2,1)$. On the other hand, $\lambda^2(G) = 6.25 > 5 = 1+4$ is achieved by $(2.5, 0)$ and $(0, 2.5)$.

We start with a result showing how the parameters $h(\mathcal{R})$, $\lambda(\mathcal{R})$, and $\delta(\mathcal{R})$ relate to each other.

Proposition 4.4. *Let $\mathcal{R}$ be a G-system. We have*

$$h(\mathcal{R}) \leq \min\left\{\frac{\lambda(\mathcal{R})}{k}, \delta(\mathcal{R})\right\}.$$

Proof. We first show that $h(\mathcal{R}) \leq \lambda(\mathcal{R})/k$. Suppose that $h(\mathcal{R}) > \lambda(\mathcal{R})/k$. By definition, we have $(h(\mathcal{R}), \ldots, h(\mathcal{R})) \in \Lambda(\mathcal{R})$. Therefore

$$\lambda(\mathcal{R}) \geq h(\mathcal{R})k > \frac{\lambda(\mathcal{R})}{k}k = \lambda(\mathcal{R}),$$

which is a contradiction. For the second part of the proof, assume that $h(\mathcal{R}) > \delta(\mathcal{R})$. Then, by the definition of $\delta(\mathcal{R})$ there must exist at least one element of the set $\{h(\mathcal{R})e_i \mid 1 \le i \le k\} \subseteq \mathbb{F}_q^k$ that does not belong to $\Lambda(\mathcal{R})$. This contradicts the definition of $h(\mathcal{R})$. ∎

Remark 4.5. The bound of Proposition 4.4 is met with equality for the service rate region depicted in Figure 1(a). However, the bound is not sharp in general. Consider, for instance, the service rate region $\Lambda(G)$ for

$$G = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \in \mathbb{F}_2^{3\times 4}.$$

Note that $(0,0,1), (0,1,0), (1,0,1), (2,1,0) \in \Lambda(G)$. It can be shown that $h(G) = 0.5$, $\lambda(G) = 3$, and $\delta(G) = 1$.

In the next example, we show that the values $\delta(\mathcal{R})$ and $\lambda(\mathcal{R})/k$ are not comparable in general, showing that taking the minimum in the bound of Proposition 4.4 is indeed needed.

Example 4.6. For the service rate region of Example 3.6, we have $\delta(G) = 2.5 > 1.5 = \lambda(G)/2$. However, for the service rate region of Figure 1(b) we have $\delta(G) = 1 < 2 = \lambda(G)/2$.

The quantity $\lambda^2(\mathcal{R})$ has a precise geometric significance; it gives the smallest sphere wedge that contains the service rate region. To illustrate how $\lambda^2(\mathcal{R})$ relates to the other fundamental parameters, we will use an argument based on the Bhatia–Davis inequality [7] from statistics. Note that the following bound is sharp for the G-system of Example 4.18 below.

Theorem 4.7. *Let $\mathcal{R}$ be a G-system. We have*

$$\lambda^2(\mathcal{R}) \le \frac{k-1}{k}\lambda^*(\mathcal{R})\lambda(\mathcal{R}) + \frac{(\lambda(\mathcal{R}))^2}{k}.$$

Proof. Let $\hat{\lambda} \in \Lambda(\mathcal{R})$ achieve $\lambda^2(\mathcal{R})$. We apply the Bhatia–Davis inequality [7] to the coordinates of $\hat{\lambda}$, obtaining

$$\begin{aligned}
\frac{1}{k}\sum_{i=1}^{k}\hat{\lambda}_i^2 &\le \left(\max_{1\le i\le k}\{\hat{\lambda}_i\} - \frac{1}{k}\sum_{i=1}^{k}\hat{\lambda}_i\right)\left(\frac{1}{k}\sum_{i=1}^{k}\hat{\lambda}_i - \min_{1\le i\le k}\{\hat{\lambda}_i\}\right) + \frac{1}{k^2}\left(\sum_{i=1}^{k}\hat{\lambda}_i\right)^2 \\
&\le \left(\frac{k-1}{k}\max_{1\le i\le k}\{\hat{\lambda}_i\}\right)\left(\frac{1}{k}\sum_{i=1}^{k}\hat{\lambda}_i\right) + \frac{1}{k^2}\left(\sum_{i=1}^{k}\hat{\lambda}_i\right)^2 \\
&\le \left(\frac{k-1}{k}\lambda^*(\mathcal{R})\right)\left(\frac{1}{k}\lambda(\mathcal{R})\right) + \frac{1}{k^2}\lambda(\mathcal{R})^2.
\end{aligned}$$

Since $\hat{\lambda}$ achieves $\lambda^2(\mathcal{R})$ by assumption, we can rewrite the inequality we just obtained as follows:

$$\frac{1}{k}\lambda^2(\mathcal{R}) \le \frac{k-1}{k^2}\lambda^*(\mathcal{R})\lambda(\mathcal{R}) + \frac{1}{k^2}(\lambda(\mathcal{R}))^2.$$

Multiplying both sides by k gives the desired result. ∎

Another natural parameter of the service rate region is its volume. Recall that the volume of a convex polytope $\mathcal{P}$ is the Lebesgue measure [20] of its interior, which we denote by $\mathrm{vol}(\mathcal{P})$.

Distributed service systems strive to support the data download of simultaneous users whose numbers and interests vary over time. The larger the service rate region volume, the larger the number of different user-number configurations the system can serve.

Computing the volume of a polytope is a difficult task in general [13]. However, some cases are relevant for our purposes, where simple observations give a closed formula for the volume of the service rate region.

Proposition 4.8. *Suppose that G is a replication matrix; see page 3 for the definition. We have*

$$\mathrm{vol}(\Lambda(G)) = \prod_{i=1}^{k} |\{1 \le \nu \le n \mid \textit{the } \nu\textit{th column of } G \textit{ is a nonzero multiple of } e_i\}|.$$

Proof. It is easy to see that the service rate region $\Lambda(G)$ is a hyperrectangle in $\mathbb{R}^k$, where each edge has a length equal to the number of times the corresponding standard basis vector appears as a column in G. The $\Lambda(G)$ volume is then determined as the quantity in the statement. ∎

In Theorem 6.5 we will give a closed formula for the volume of $\Lambda(G)$, when G generates a 3-dimensional MDS code of length at least 6. The result is embedded in section 6, which is devoted to the service rate region of systematic MDS codes.

We now compute the volume of the allocation polytope of a replication system, showing in particular that the volume of the allocation polytope does not determine the volume of the service rate region. Intuitively, this follows from the fact that the volume of the allocation polytope is multilinear in the coordinates corresponding to the same object, whereas the volume of the service rate polytope is linear in their sum. We introduce a class of polytopes that will be used later in section 5.

Definition 4.9. *A polytope of the form $\mathcal{P} = \{x \in [0,1]^m \mid yx^\top \le n\} \subseteq \mathbb{R}^m$, where n and m are positive integers and $y \in \mathbb{R}^m_{\ge 0}$ is a vector, is called a relaxed knapsack polytope in $\mathbb{R}^m$.*

The volume of a relaxed knapsack polytope as in Definition 4.9 is known to be

$$\mathrm{vol}(\mathcal{P}) = \frac{1}{m! \prod_{i=1}^m y_i} \sum_{x \in \{0,1\}^m \cap \mathcal{P}} (-1)^{\mathrm{wt}(x)} g(x)^m, \tag{4.1}$$

where $\mathrm{wt}(x)$ is the number of nonzero entries of x and $g(x) = n - \sum_{i=1}^m y_i x_i$ for all $x \in \mathbb{R}^m$; see, e.g., [6]. Using the above formula for the volume and some elementary generating functions theory, we compute the volume of the allocation polytope of a replication matrix.

Proposition 4.10. *Suppose that G is a replication matrix. We have* $\mathrm{vol}(\mathcal{A}(G)) = 1$.

Proof. It is not hard to see that $\mathcal{A}(G) = \{x \in [0,1]^n \mid x_1 + \cdots + x_n \le n\}$, which is a relaxed knapsack polytope obtained for $m = n$ and $y = 1^n = (1, \ldots, 1)$. Therefore, using (4.1) and denoting by $[x^n]S(x)$ the coefficient of x^n in a power series $S(x)$, we compute

$$\begin{aligned}\mathrm{vol}(\mathcal{A}(G)) &= \frac{1}{n!} \sum_{x \in \{0,1\}^n} (-1)^{\mathrm{wt}(x)} (n - (x_1 + \ldots + x_n))^n = \frac{1}{n!} \sum_{x \in \{0,1\}^n} (-1)^{\mathrm{wt}(x)} (n - \mathrm{wt}(x))^n \\ &= \frac{1}{n!} \sum_{i=0}^{n} (-1)^i (n-i)^n \binom{n}{i} = \sum_{i=0}^{n} (-1)^i \binom{n}{i} [x^n] e^{(n-i)x} = [x^n](e^x - 1)^n = 1,\end{aligned}$$

where all passages easily follow from binomial theorem and the Taylor expansion of the exponential function. ∎

Note that even though the volume of the allocation polytope is the same for every replication matrix, the volume of the service rate region is not a constant; see Proposition 4.8.

Throughout this section, we focus on the connection between the parameters of $\Lambda(G)$ and those of the error-correcting code generated by G; see Appendix B for some coding theory background.

We start by recalling the following result from [2], whose statement relies on interpreting the columns of G as points of the finite projective space $\mathrm{PG}(k-1,q)$; see [31] for a general reference. This can be done because, as stated in Notation 1.1, none of the columns of G is the zero vector.

Proposition 4.11. *Let $\lambda \in \Lambda(G)$ and let $I \subseteq \{1,\dots,k\}$ be an index set. Let H be a hyperplane of $\mathrm{PG}(k-1,q)$ not containing any of the standard basis vectors e_i, for $i \in I$, and let S denote the multiset of columns of G in $\mathrm{PG}(k-1,q)$. We have*

$$\sum_{i\in I} \lambda_i \le |S \setminus H|,$$

where $S \setminus H$ is the multiset of points obtained from S after removing all the points contained in H, counted with their multiplicity.

The following lemma is well known and can be shown by considering the columns of G as a multiset of points in $\mathrm{PG}(k-1,q)$, which are not all contained in a hyperplane since G has rank k by assumption.

Lemma 4.12. *Let S be the multiset of the columns of G, viewed as projective points in $\mathrm{PG}(k-1,q)$. Let d be the minimum distance of the code generated by G. Then every hyperplane of $\mathrm{PG}(k-1,q)$ contains at most $n-d$ points of S, and there exists a hyperplane of $\mathrm{PG}(k-1,q)$ which contains exactly $n-d$ points of S.*

We can also apply Proposition 4.11 to show a connection between the minimum distance of the code generated by G and the largest simplex contained in the service rate region.

Corollary 4.13. *Let d denote the minimum distance of the code generated by G. We have*

$$\lceil \delta(G) \rceil \le d.$$

Proof. Let $i \in \{1,\dots,k\}$ be fixed and let $H \subseteq \mathrm{PG}(k-1,q)$ be a hyperplane that does not contain e_i. By applying Proposition 4.11 with $I = \{i\}$ and $\lambda = \delta(G)e_i \in \Lambda(G)$, we have

$$\delta(G) \le |S \setminus H| = n - |S \cap H|,$$

where S is the multiset of columns of G. We can follow the same reasoning for every i. Since every hyperplane H does not contain some e_i,

$$d = n - \max\{|S \cap H| : H \subseteq \mathrm{PG}(k-1,q),\ H \text{ hyperplane}\},$$

and $\delta(G) \le |S \setminus H|$, we conclude that $\lceil \delta(G) \rceil \le d$. ∎

The bound of Corollary 4.13 is met with equality by some matrices G, as the following example illustrates.

Example 4.14. Let

$$G = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \in \mathbb{F}_2^{3\times 4}.$$

It can be easily seen that $2e_i \in \Lambda(G)$ for $i \in \{1,2,3\}$, and therefore $\delta(G) = 2$.

In the last part of this section, inspired by the coding theory literature, we introduce the notion of *availability* for the matrix G. We then describe the role this notion plays in shaping the geometry of the service rate region.

Definition 4.15. *Suppose that G is systematic. We say that G has* availability $t \in \mathbb{Z}_{\geq 0}$ *if $\mathcal{R}_i^{\text{all}}(G)$ contains $t+1$ pairwise disjoint sets for all $i \in \{1,\dots,k\}$.*

The following result easily follows from the definitions.

Proposition 4.16. *Suppose that G is systematic and has availability t. Then $(t+1)e_i \in \Lambda(G)$ for all $i \in \{1,\dots,k\}$. In particular, $\lfloor \delta(G) \rfloor \geq t+1$.*

By combining Corollary 4.13 with Proposition 4.16 we obtain the following result.

Corollary 4.17. *Suppose that G is systematic and has availability t. Let d denote the minimum distance of the code generated by G. We have $d \geq t+1$.*

We conclude this section with an example where Proposition 4.16 and Corollary 4.17 are sharp.

Example 4.18 (the simplex code). Let

$$G = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix} \in \mathbb{F}_2^{3\times 7}.$$

Then G has availability 3 and Proposition 4.16 is sharp in this case. Note that G is the generator matrix of one of the best known error-correcting codes, namely the simplex code; see, e.g., [21].

5. Outer bounds.

In this section, we derive outer bounds for the service rate region $\Lambda(G)$ as bounding polytopes $\mathcal{P} \supseteq \Lambda(G)$. We apply methods from coding theory and optimization, dedicating a subsection to each of the two approaches. We illustrate how to apply the bounds with examples and comment on their sharpness.

5.1. Coding theory approach.

We start with a simple result that can be easily obtained by summing the inequalities that define the allocation polytope, namely the constraints in (1.2) for $1 \leq \nu \leq n$.

Lemma 5.1 (total capacity bound). *Let $\mathcal{R}$ be a G-system and let $\{\lambda_{i,R}\}$ be a feasible allocation for $(\mathcal{R},\mu)$; see Definition 1.4. We have*

$$\sum_{i=1}^{k} \sum_{R \in \mathcal{R}_i} |R| \lambda_{i,R} \leq \mu n. \tag{5.1}$$

The following result links the size of the recovery sets of a G-system to the parameters of the (dual of the) error-correcting code generated by G.

Proposition 5.2. *Suppose that G is systematic. Let $d^\perp$ denote the minimum distance of the dual of the code generated by G. For all $i \in \{1, \dots, k\}$ and $R \in \mathcal{R}_i^{\text{all}}(G)$ we have $R = \{i\}$ or $|R| \geq d^\perp - 1$.*

We now establish the first outer bound of this section.

Theorem 5.3 (dual distance bound). *Suppose that G is systematic. Let $d^\perp$ denote the minimum distance of the dual of the code generated by G. If $(\lambda_1, \dots, \lambda_k) \in \Lambda(G)$, then*

$$\sum_{i=1}^{k} (\min\{\lambda_i, 1\} + (d^\perp - 1) \max\{0, \lambda_i - 1\}) \leq n.$$

Proof. Let $(\lambda_1, \dots, \lambda_k) \in \Lambda(G)$ and let $\{\lambda_{i,R}\}$ be a feasible allocation for $(\mathcal{R}, 1)$. By Proposition 5.2 we have $|R| \geq d^\perp - 1$ for every $i \in \{1, \dots, k\}$ and $R \in \mathcal{R}_i^{\text{all}}(G)$ with $R \neq \{i\}$. We can therefore rewrite the left-hand side of (5.1) as follows:

$$\begin{aligned}
\sum_{i=1}^{k} \lambda_{i,\{i\}} + \sum_{i=1}^{k} \sum_{\substack{R \in \mathcal{R}_i^{\text{all}}(G) \\ R \neq \{i\}}} |R| \lambda_{i,R} &\geq \sum_{i=1}^{k} \lambda_{i,\{i\}} + (d^\perp - 1) \sum_{i=1}^{k} \sum_{\substack{R \in \mathcal{R}_i^{\text{all}}(G) \\ R \neq \{i\}}} \lambda_{i,R} \\
&= \sum_{i=1}^{k} \lambda_{i,\{i\}} + (d^\perp - 1) \sum_{i=1}^{k} \left(\lambda_i - \lambda_{i,\{i\}}\right) \\
&= (d^\perp - 1) \sum_{i=1}^{k} \lambda_i - (d^\perp - 2) \sum_{i=1}^{k} \lambda_{i,\{i\}}. \qquad (5.2)
\end{aligned}$$

Since G has no all-zero column, we have $d^\perp \geq 2$. Therefore, using the fact that $\lambda_{i,\{i\}} \leq \min\{\lambda_i, 1\}$ for all i, we can further say that the right-hand side of (5.2) is at least

$$(d^\perp - 1) \sum_{i=1}^{k} \lambda_i - (d^\perp - 2) \sum_{i=1}^{k} \min\{\lambda_i, 1\} = \sum_{i=1}^{k} (\min\{\lambda_i, 1\} + (d^\perp - 1) \max\{0, \lambda_i - 1\}),$$

which, combined with (5.1), gives the statement. ∎

Remark 5.4. It follows from [4] that the dual distance bound of Theorem 5.3 is sharp if G is a systematic MDS matrix and $n \geq 2k$; see Appendix B for the definition of an MDS matrix. The bound can be sharp also for systematic matrices $G \in \mathbb{F}_q^{k \times n}$ that generate an MDS code and have $n < 2k$. This is the case of the matrix G of Example 4.14.

It turns out that Theorem 5.3 is not particularly effective for systems that mainly implement replication, i.e., for matrices G that are very similar to a replication matrix. We obtain the following result by considering the number of systematic nodes for each object. Since the proof is similar to the one of Theorem 5.3, we omit it here.

Theorem 5.5. *Suppose that G is systematic and let s_i denote the number of systematic nodes for the ith object, for $i \in \{1,\dots,k\}$. If $(\lambda_1,\dots,\lambda_k) \in \Lambda(G)$, then*

$$\sum_{i=1}^{k} (\min\{\lambda_i, s_i\} + 2\max\{0, \lambda_i - s_i\}) \le n.$$

The outer bounds Theorems 5.3 and 5.5 are not generally comparable, as the following example illustrates.

Example 5.6. An example where Theorem 5.3 outperforms Theorem 5.5 is given by the region in Example 4.14. By Remark 5.4, the bound of Theorem 5.3 gives the exact service rate region. This automatically outperforms the bound of Theorem 5.5 as $(d^\perp - 1) = 3 > 2$. Now consider the service rate region of Example 1.11 depicted in Figure 1(b). The bounding polytopes given by Theorems 5.3 and 5.5 are depicted in Figure 3, showing that Theorem 5.5 outperforms Theorem 5.3 in that case.

The next result is a hybrid between Theorems 5.3 and 5.5, in the sense that it takes into account *both* the minimum distance of the dual of the code generated by G and the number of systematic nodes.

Theorem 5.7. *Let $d^\perp$ be the minimum distance of the dual of the code generated by G and let s_i denote the number of systematic nodes for the ith object, for $i \in \{1,\dots,k\}$. For all $(\lambda_1,\dots,\lambda_k) \in \Lambda(G)$ we have*

$$\sum_{\substack{i\in\{1,\dots,k\}\\ s_i\neq 0}} (\min\{s_i, \lambda_i\} + \max\{2, d^\perp - 1\}\max\{0, \lambda_i - s_i\}) + \sum_{\substack{i\in\{1,\dots,k\}\\ s_i = 0}} 2\lambda_i \le n.$$

Proof. Let $(\lambda_1,\dots,\lambda_k) \in \Lambda(G)$ and let $\{\lambda_{i,R}\}$ be a corresponding feasible allocation. Write $\mathcal{R}$ for $\mathcal{R}^{\min}(G)$. By Proposition 5.2 and the fact that $d^\perp = 2$ if there exist two columns of G that are linearly dependent, we obtain

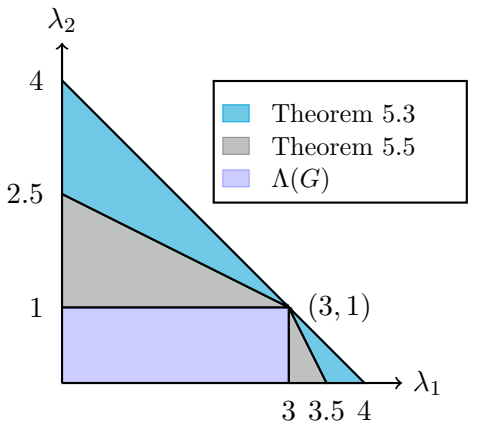


Figure 3. *The service rate region of Example 1.11, Figure 1(b), is an example where the bound of Theorem 5.5 gives a better approximation than the bound of Theorem 5.3.*

$$\sum_{i\in\{1,\dots,k\}}\sum_{R\in\mathcal{R}}|R|\lambda_{i,R}\geq\sum_{\substack{i\in\{1,\dots,k\}\\ s_i\neq 0}}\left(\sum_{\substack{R\in\mathcal{R}\\ |R|=1}}\lambda_{i,R}+\max\{2,d^\perp-1\}\sum_{\substack{R\in\mathcal{R}\\ |R|\neq 1}}\lambda_{i,R}\right)+\sum_{\substack{i\in\{1,\dots,k\}\\ s_i=0}}\sum_{R\in\mathcal{R}}2\lambda_{i,R}.$$

Therefore,

$$\begin{aligned}
&\sum_{i\in\{1,\dots,k\}}\sum_{R\in\mathcal{R}}|R|\lambda_{i,R}\\
&\quad\geq\sum_{\substack{i\in\{1,\dots,k\}\\ s_i\neq 0}}\sum_{\substack{R\in\mathcal{R}\\ |R|=1}}\lambda_{i,R}+\sum_{\substack{i\in\{1,\dots,k\}\\ s_i\neq 0}}\left(\max\{2,d^\perp-1\}\left(\lambda_i-\sum_{\substack{R\in\mathcal{R}\\ |R|=1}}\lambda_{i,R}\right)\right)+\sum_{\substack{i\in\{1,\dots,k\}\\ s_i=0}}2\lambda_i\\
&\quad\geq\sum_{\substack{i\in\{1,\dots,k\}\\ s_i\neq 0}}\max\{2,d^\perp-1\}\lambda_i-(\max\{2,d^\perp-1\}-1)\sum_{\substack{i\in\{1,\dots,k\}\\ s_i\neq 0}}\min\{s_i,\lambda_i\}\\
&\qquad+\sum_{\substack{i\in\{1,\dots,k\}\\ s_i=0}}2\lambda_i\\
&\quad=\sum_{\substack{i\in\{1,\dots,k\}\\ s_i\neq 0}}(\min\{s_i,\lambda_i\}+\max\{2,d^\perp-1\}(\lambda_i-\min\{s_i,\lambda_i\}))+\sum_{\substack{i\in\{1,\dots,k\}\\ s_i=0}}2\lambda_i\\
&\quad=\sum_{\substack{i\in\{1,\dots,k\}\\ s_i\neq 0}}(\min\{s_i,\lambda_i\}+\max\{2,d^\perp-1\}\max\{0,\lambda_i-s_i\})+\sum_{\substack{i\in\{1,\dots,k\}\\ s_i=0}}2\lambda_i,
\end{aligned}$$

where the first inequality follows from Definition 1.4, and the second follows from the inequality

$$\sum_{\substack{R\in\mathcal{R}\\ |R|=1}}\lambda_{i,R}\leq\min\{s_i,\lambda_i\}.$$

■

It is interesting to note that if $d^\perp-1\geq 2$ (and hence $s_i=1$ for all $i\in\{1,\dots,k\}$), then Theorem 5.7 gives Theorem 5.3. Similarly, if $s_i\neq 0$ for all $i\in\{1,\dots,k\}$ and $s_i\geq 2$ for at least one i, then Theorem 5.7 becomes Theorem 5.5.

Lemma 5.1 suggests that the variety of sizes of the recovery sets plays an important role in shaping the service rate region. By taking into account the indices i for which the recovery sets all have the same size, we obtained the following result. Note that this exactly measures the contribution of the indices for which the recovery sets have the same size and thus improves upon Theorem 5.7. The proof is a simple extension of Theorem 5.7, and we omit it here.

Theorem 5.8. *Let $d^\perp$ be the minimum distance of the dual of the code generated by G and let s_i denote the number of systematic nodes for the ith object, for $1\in\{1,\dots,k\}$. Let*

$$\mu_i = \frac{1}{|\mathcal{R}_i^{\min}(G)| - s_i} \sum_{\substack{R \in \mathcal{R}_i^{\min}(G) \\ |R| \neq 1}} |R| \quad \text{for } i \in \{1, \dots, k\},$$

$$J = \{i \in \{1, \dots, k\} \mid \text{all } R \in \mathcal{R}_i^{\min}(G) \text{ with } |R| \neq 1 \text{ have the same cardinality}\}.$$

Then for all $\lambda \in \Lambda(G)$ *we have*

$$n \geq \sum_{\substack{i \in \{1,\dots,k\} \\ s_i \neq 0,\, i \notin J}} (\min\{s_i, \lambda_i\} + \max\{2, d^\perp - 1\} \max\{0, \lambda_i - s_i\}) + \sum_{\substack{i \in \{1,\dots,k\} \\ s_i = 0,\, i \notin J}} 2\lambda_i$$
$$+ \sum_{\substack{i \in \{1,\dots,k\} \\ s_i = 0,\, i \in J}} \mu_i \lambda_i + \sum_{\substack{i \in \{1,\dots,k\} \\ s_i \neq 0,\ |\mathcal{R}_i^{\min}(G)| = s_i}} s_i + \sum_{\substack{i \in \{1,\dots,k\} \\ s_i \neq 0,\, i \in J}} (\mu_i \lambda_i - (1 - \mu_i) \min\{s_i, \lambda_i\}).$$

In the next example, we show that Theorem 5.8 can be sharper than Theorem 5.7 for some service rate regions.

Example 5.9. Let $k = 3$, $n = 6$, $q = 3$, and

$$G := \begin{pmatrix} 0 & 1 & 1 & 2 & 1 & 2 \\ 1 & 2 & 2 & 2 & 1 & 1 \\ 0 & 0 & 0 & 1 & 2 & 2 \end{pmatrix} \in \mathbb{F}_3^{3 \times 6}.$$

Following the notation of Theorem 5.8 we have $d^\perp = 2$, $(\mu_1, \mu_2, \mu_3) = (2, 3, 11/4)$, $(s_1, s_2, s_3) = (0, 1, 0)$, and $J = \{1, 2\}$. Figure 4 depicts the service rate region $\Lambda(G)$ and the outer bounds given by Theorems 5.7 and 5.8.

5.2. Optimization approach. In this subsection, we use the theory of knapsack polytopes (recall Definition 4.9) to derive an outer bound for the allocation and service rate region polytopes and the bound corollaries. Most notably, we obtain upper bounds for the quantities $\sum_{i \in I} \lambda_i$, for any given index set $I \subseteq \{1, \dots, k\}$. These quantities are of interest in practice

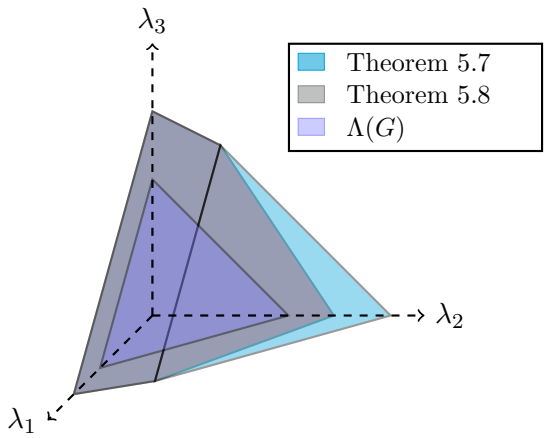


Figure 4. *Service rate region and outer bounds for Example* 5.9.

because they correspond to the cumulative numbers of users interested in some (sub)sets of stored objects. Observe that when $I = \{1, \dots, k\}$, then $\sum_{i \in I} \lambda_i$ represents the total number of users in the system. In this subsection, we also illustrate how to apply our bounds in some examples and plot the output.

Notation 5.10. *Let $\mathcal{R}$ be a G-system and $m(\mathcal{R}) = |\mathcal{R}_1| + \cdots + |\mathcal{R}_k|$. We define the integer vector*

$$y(\mathcal{R}) = (|R| : i \in \{1, \dots, k\}, R \in \mathcal{R}_i) \in \mathbb{Z}_{\geq 0}^{m(\mathcal{R})},$$

where we take the same order as in Definition 2.1. Note that all the entries of $y(\mathcal{R})$ are positive since recovery sets are nonempty by definition.

We can now state the main result of this section, which gives an infinite number of half-spaces that contain the allocation polytope, one for each vector $c \in \mathbb{R}^{m(\mathcal{R})}$.

Theorem 5.11. *Let $\mathcal{R}$ be a G-system, $m = m(\mathcal{R})$, and let $c \in \mathbb{R}^m$. Define $y = y(\mathcal{R})$ and let $\pi : \{1, \dots, m\} \to \{1, \dots, m\}$ be any permutation such that*

$$\frac{c_{\pi(1)}}{y_{\pi(1)}} \geq \cdots \geq \frac{c_{\pi(m)}}{y_{\pi(m)}}.$$

Define $J = \{j \mid y_{\pi(1)} + \cdots + y_{\pi(j)} > n\}$. If $J = \emptyset$, then let $r = m + 1$, $\sigma = 0$, and $\pi(m+1) = 0$. If $J \neq \emptyset$, then let $r = \min(J)$ and $\sigma = (n - \sum_{j=1}^{r-1} y_{\pi(j)})/y_{\pi(r)}$. Then for any $x \in \mathcal{A}(\mathcal{R})$ we have

$$cx^\top \leq \left(\sum_{j=1}^{r-1} c_{\pi(j)} \right) + c_{\pi(r)}\, \sigma, \quad \text{where } c_{\pi(m+1)} = 0. \tag{5.3}$$

Before proving the theorem, we state an immediate consequence for the max-sum capacity of the service rate region. The result is obtained by taking $c = (1, \dots, 1)$, allowing us to use a more efficient notation.

Corollary 5.12. *Let $\mathcal{R}$ be any G-system with the property that $\Lambda(\mathcal{R}) = \Lambda(G)$. Let $y = y(\mathcal{R})$ and reorder its components nondecreasingly obtaining a vector $\hat{y}$. Suppose $\hat{y}_1 + \cdots + \hat{y}_m > n$, where $m = m(\mathcal{R})$, and let $r = \min\{j \mid \hat{y}_1 + \cdots + \hat{y}_j > n\}$. We have*

$$\lambda(G) \leq r - 1 + \frac{n - \sum_{j=1}^{r-1} \hat{y}_j}{\hat{y}_r}.$$

We give an example illustrating how to apply Corollary 5.12.

Example 5.13. Let G be as in Example 4.14, with $n = 4$, and $\mathcal{R} = \mathcal{R}^{\min}(G)$. We have $y = y(\mathcal{R}) = (1, 3, 1, 3, 1, 3)$. As in Corollary 5.12, we construct $\hat{y} = (1, 1, 1, 3, 3, 3)$ and obtain $\lambda(G) \leq 10/3$. It can be checked that $\lambda(G) = 3$.

Proof of Theorem 5.11. Let $\mathcal{P} = \{x \in [0,1]^{m(\mathcal{R})} \mid y(\mathcal{R})x^\top \leq n\}$, which is a relaxed knapsack polytope. Let $\beta = \max\{cx^\top \mid x \in \mathcal{P}\}$. We have the inclusion $\mathcal{A}(\mathcal{R}) \subseteq \mathcal{P}$; hence $\max\{cx^\top \mid x \in \mathcal{A}(\mathcal{R})\} \leq \beta$. We apply a classical result by Dantzig [9] (the case where $J = \emptyset$ requires a separate treatment, but we omit it here), which states that a point $\hat{x} \in \mathcal{P}$ attaining the maximum β is given by

$$\hat{x}_j = \begin{cases} 1 & \text{if } 1 \le j \le r-1, \\ \dfrac{n - \sum_{j=1}^{r-1} y_{\pi(j)}}{y_{\pi(r)}} & \text{if } j = r, \\ 0 & \text{otherwise.} \end{cases}$$

The result follows by computing $\mu = c\hat{x}^\top$. ∎

As another corollary of Theorem 5.11, we obtain a result that gives an infinite number of half-spaces in which the service rate region $\Lambda(\mathcal{R})$ is contained. Each half-space is obtained by choosing a different vector b in the statement.

Corollary 5.14. *Let $\mathcal{R}$ be a G-system and let $b \in \mathbb{R}^k$. Let $m = m(\mathcal{R})$, $m_0 = 1$, and $m_i = |\mathcal{R}_i|$ for all $i \in \{1, \dots, k\}$. Define $c \in \mathbb{R}^{m(\mathcal{R})}$ by setting $c_j = b_i$ whenever $m_{i-1} + 1 \le j \le m_i$. Construct a permutation π and define r, σ, and $\pi(m+1)$ if necessary, as in Theorem 5.11. Then for all $\lambda \in \Lambda(\mathcal{R})$ we have*

$$b\lambda^\top \le \left(\sum_{j=1}^{r-1} c_{\pi(j)} \right) + c_{\pi(r)}\, \sigma.$$

By specializing the previous result to vectors $b \in \{0,1\}^k$ one can obtain upper bounds for partial sums of the form $\sum_{i \in I} \lambda_i$, where $I \subseteq \{1, \dots, k\}$ and $\lambda \in \Lambda(\mathcal{R})$. In particular, one can obtain an upper bound for the max-sum capacity $\lambda(G)$. We conclude this section by illustrating how Corollary 5.14 can be applied and the type of results it gives.

Example 5.15. Consider the service rate region of Example 1.11, depicted in Figure 1(b). By applying Corollary 5.14 for all $b \in \{0,1\}^3$ we obtain the bounding polytope for the service rate region, and we depict it in Figure 5 as well as with $\Lambda(G)$.

The following example shows that applying Corollary 5.14 not only with 0-1 vectors can give a strictly better bound than only applying it with 0-1 vectors.

Example 5.16. Let

$$G = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 2 \end{pmatrix} \in \mathbb{F}_3^{2\times 8}.$$

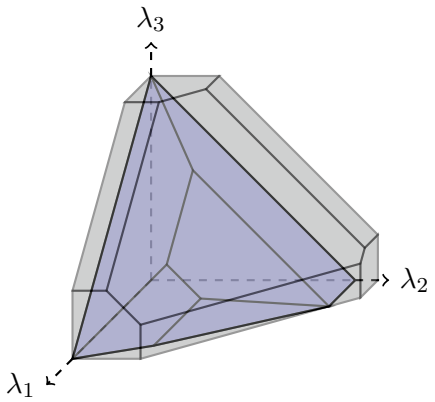


Figure 5. *Service rate region and bounding polytope for Example 5.15.*

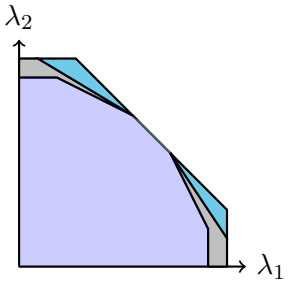


Figure 6. *The service rate region and the bounding polytopes for Example 5.16.*

The service rate region $\Lambda(G)$ is the purple region in Figure 6. By applying Corollary 5.14 for all $b \in \{0,1\}^2$, one gets the light blue region in Figure 6. For a better approximation of the service rate region, also depicted in Figure 6, we can use Corollary 5.14. For example, by applying said corollary with $b = (3,2)$ and $b = (3,5)$, in addition to the 0-1 vectors $b \in \{0,1\}^2$, one gets the gray region in Figure 6.

6. Systematic MDS codes. This section is entirely devoted to the service rate region $\Lambda(G)$, when G is an MDS matrix; see Appendix B for the definition of MDS matrix. We focus on the volumes of these service rate regions for $k \in \{2,3\}$ and on their max-sum capacities.

Recall that in the case where G is a systematic MDS matrix and $n \geq 2k$, the service rate region $\Lambda(G)$ is known and given by the set

$$\left\{ (\lambda_1, \dots, \lambda_k) \in \mathbb{R}^k_{\geq 0} \;\middle|\; \sum_{i=1}^{k} \Big(\min\{\lambda_i, 1\} + k \cdot \max\{0, \lambda_i - 1\} \Big) \leq n \right\}; \tag{6.1}$$

see Remark 5.4. The description in (6.1) is however inconvenient for computing the volume of $\Lambda(G)$, which is one of the goals of this section. Therefore, our first move is deriving a more convenient description.

Notation 6.1. *Given a vector $\lambda \in \mathbb{R}^k$, let $\chi(\lambda) \in \mathbb{Z}^k$ be the vector with $\chi(\lambda)_i = 1$ if $\lambda_i < 1$ and $\chi(\lambda)_i = k$ if $\lambda_i \geq 1$. Moreover, we let*

$$\lambda_{<1} = \{ i \in \{1, \dots, k\} \mid \lambda_i < 1 \}.$$

The following lemma gives a different representation of the service rate region of a systematic MDS matrix G with $n \geq 2k$. In its statement, we use Notation 6.1.

Lemma 6.2. *Suppose G is a systematic MDS matrix and $n \geq 2k$. We have*

$$\Lambda(G) = \left\{ \lambda \in \mathbb{R}^k_{\geq 0} \;\middle|\; \chi(\lambda)\, \lambda^\top \leq n + (k-1)(k - |\lambda_{<1}|) \right\}. \tag{6.2}$$

Proof. Let $\lambda \in \Lambda(G)$. By (6.1), we have $\sum_{i=1}^{k} \Big(\min\{\lambda_i, 1\} + k \cdot \max\{0, \lambda_i - 1\} \Big) \leq n$. That is,

$$\sum_{i \in \lambda_{<1}} \lambda_i + (k - |\lambda_{<1}|) + \left(\sum_{i \in \{1,\dots,k\} \setminus \lambda_{<1}} k\lambda_i \right) - k(k - |\lambda_{<1}|) \leq n.$$

The latter inequality can be rewritten as $\chi(\lambda)\cdot\lambda^\top \le n+(k-1)(k-|\lambda_{<1}|)$. This shows the inclusion $\subseteq$ in (6.2). The other inclusion follows by reversing all the passages, and we omit the details. ∎

We can now compute the volume of the service rate region of an MDS matrix for $k\in\{2,3\}$ and $n\ge 2k$. We start with the case $k=2$.

Theorem 6.3. *Let $G\in\mathbb{F}_q^{2\times n}$ be a systematic MDS matrix. Suppose $n\ge 4$. Then we have* $\mathrm{vol}(\Lambda(G))=\frac{n^2+4n}{8}$.

Proof. By Lemma 6.2, $\Lambda(G)$ is defined by the following five equations:

$$\lambda_1+\lambda_2\le\frac{n+2}{2},\qquad 2\lambda_1+\lambda_2\le n+1,\qquad \lambda_1+2\lambda_2\le n+1,\qquad \lambda_1\ge 0,\qquad \lambda_2\ge 0.$$

It is not hard to check that the vertices of $\Lambda(G)$ are the points

$$(0,0),\qquad \left(1,\frac{n}{2}\right),\qquad \left(\frac{n}{2},1\right),\qquad \left(\frac{n+1}{2},0\right),\qquad \left(0,\frac{n+1}{2}\right).$$

The volume (i.e., the area) can now be computed using elementary methods. ∎

We can compare Theorem 6.3 with a replication system generated by a matrix with the same parameters.

Proposition 6.4. *Suppose that $G\in\mathbb{F}_q^{k\times n}$ is a replication matrix. We have*

$$n-k+1\le\mathrm{vol}(\Lambda(G))\le\left\lfloor\left(\frac{n}{k}\right)^k\right\rfloor.$$

The lower bound can be attained by some G and the upper bound can be attained by some G if $k=2$.

Proof. Let j_i denote the number of columns of G multiples of the standard basis vector e_i, for $i\in\{1,\dots,k\}$. Each j_i is a positive integer since G is full rank. By Proposition 4.8, the volume of $\Lambda(G)$ is $\prod_{i=1}^k j_i$. We get the desired upper bound by the arithmetic versus geometric mean inequality. The lower bound can be attained by taking $j_1=n-k+1$ and $j_i=1$ for $i\in\{2,\dots,k\}$. When $k=2$, the upper bound can be attained by taking $j_1\in\{\lfloor n/2\rfloor,\lceil n/2\rceil\}$ and $j_2=n-j_1$. ∎

Note that Proposition 6.4 shows that one can find a replication matrix $G\in\mathbb{F}_q^{2\times n}$ for $n\ge 5$ whose service rate region's volume is strictly larger than the volume of an MDS matrix of the same size.

We now turn to the case $k=3$ and $n\ge 6$, computing the volume of the service rate region corresponding to an MDS matrix $G\in\mathbb{F}_q^{3\times n}$. The computation is more involved than in the 2-dimensional case.

Theorem 6.5. *Let $G\in\mathbb{F}_q^{3\times n}$ be a systematic MDS matrix and suppose $n\ge 6$. We have*

$$\mathrm{vol}(\Lambda(G))=\frac{n^3+18n^2+54n-18}{162}.$$

Proof. Define the function $f(z)=\min\{z,1\}+3\max\{0,z-1\}$. Using Lemma 6.2 it can be seen that $\Lambda(G)$ is the set of 3-tuples (λ_1,λ_2,z) that satisfy the inequalities

$$\begin{cases} \lambda_1+\lambda_2 \le n-f(z), & (*)\\ 3\lambda_1+\lambda_2 \le n-f(z)+2,\\ \lambda_1+3\lambda_2 \le n-f(z)+2,\\ 3\lambda_1+3\lambda_2 \le n-f(z)+4, & (**)\\ \lambda_1,\lambda_2,z \ge 0. \end{cases}$$

Observe moreover that the maximum value z can take is $(n+2)/3$. This value can be attained by taking $\lambda_1=\lambda_2=0$ in the above system. We have

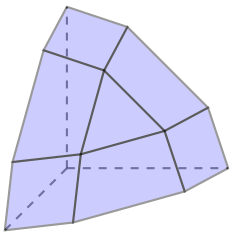
$$f(z)=\begin{cases} z & \text{if } \ 0\le z\le 1,\\ 3z-2 & \text{if } \ 1\le z. \end{cases}$$

It can be checked that $(**)$ is more restrictive than $(*)$ for $z \le n/3$, while $(*)$ is more restrictive than $(**)$ otherwise. Moreover, when $z \ge (n+1)/3$, all inequalities except for $(*)$ and the nonnegativity of λ_1, λ_2, and z can be disregarded. This tells us the shape of the "slices" of $\Lambda(G)$ for a given value of z. We summarize this discussion in Table 1 and Figure 7.

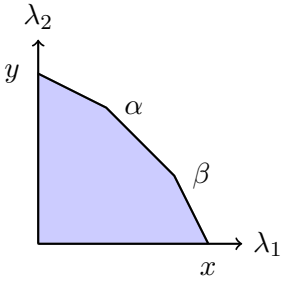
The areas of the slices can be easily computed and therefore the volume of $\Lambda(G)$ can be computed by integration over z. This approach is mathematically justified, for example, by [5, Theorem 2.7]. The desired formula follows from

Table 1

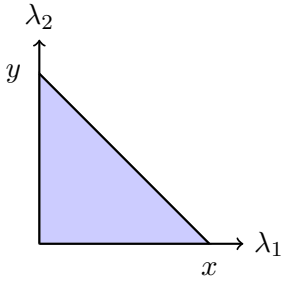
z	Figure	y	x	α	β
$0\le z\le 1$	7(b)	$(0,\frac{n-z+2}{3})$	$(0,\frac{n-z+2}{3})$	$(1,\frac{n-z+1}{3})$	$(\frac{n-z+1}{3},1)$
$1\le z\le \frac{n}{3}$	7(b)	$(0,\frac{n-3z+4}{3})$	$(0,\frac{n-3z+4}{3})$	$(1,\frac{n-3z+3}{3})$	$(\frac{n-3z+3}{3},1)$
$\frac{n}{3}\le z\le \frac{n+1}{3}$	7(b)	$(0,\frac{n-3z+4}{3})$	$(0,\frac{n-3z+4}{3})$	$(n-3z+1,1)$	$(1,n-3z+1)$
$\frac{n+1}{3}\le z\le \frac{n+2}{3}$	7(c)	$(0,n-3z+2)$	$(n-3z+2,0)$	—	—



(a) The typical service rate region of a systematic MDS matrix $G\in\mathbb{F}_q^{3\times n}$ for $n\ge 6$.



(b) The slice of the service rate region of a systematic MDS matrix $G\in\mathbb{F}_q^{3\times n}$, $n\ge 6$, for $z\le (n+1)/3$.



(c) The slice of the service rate region of a systematic MDS matrix $G\in\mathbb{F}_q^{3\times n}$, $n\ge 6$, for $(n+1)/3\le z\le (n+2)/3$.

Figure 7. *The service rate region and its slices for the proof of Theorem 6.5.*

$$\begin{aligned}\operatorname{vol}(\Lambda(G)) = &\int_0^1 \left(\frac{1}{18}z^2 - \frac{n+4}{9}z + \frac{n^2+8n+4}{18}\right) dz \\ &+ \int_1^{n/3} \left(\frac{1}{2}z^2 - \frac{n+6}{3}z + \frac{n^2+12n+24}{18}\right) dz \\ &+ \int_{n/3}^{(n+1)/3} \left(-\frac{3}{2}z^2 + (n-2)z + \frac{n^2-4n-8}{6}\right) dz \\ &+ \int_{(n+1)/3}^{(n+2)/3} \left(\frac{9}{2}z^2 - (3n+6)z + \frac{n^2+4n+4}{2}\right) dz\end{aligned}$$

and tedious but straightforward computations. ∎

Out of curiosity, we point out that Theorem 6.5 can also be derived by the well-known *triangulation method* for computing the volume of a polytope using the volume of simplices; see [8]. For the polytope of Theorem 6.5 the formalization of this approach is rather involved, which is why we proceeded by integration.

We also notice that a general lower bound for $\operatorname{vol}(\Lambda(G))$ where G is an MDS matrix can be obtained by Proposition B.3. Any k columns can be used to recover any data object. Thus, $(n/k)e_i \in \Lambda(G)$, which implies that the simplex with these vertices is contained in the service rate region. Therefore, $\operatorname{vol}(\Lambda(G)) \geq (n/k)^k/k!$.

In the second part of this section, we investigate other parameters of the service rate regions of systematic MDS matrices. We first observe that Corollary 5.12 implies the following.

Corollary 6.6. *Let $G \in \mathbb{F}_q^{k\times n}$ be a systematic MDS matrix. We have $\lambda(G) \leq k + \frac{n-k}{k}$.*

Proof. Following the notation of Corollary 5.12, we have $m = k\left(\binom{n-1}{k}+1\right)$ and $\hat{y} = (v_1, v_2)$, where $v_1 = (1,\dots,1) \in \mathbb{R}^k$ and $v_2 = (k,\dots,k) \in \mathbb{R}^{m-k}$. Moreover,

$$r = \min\{j \mid \hat{y}_1 + \cdots + \hat{y}_j > n,\ 1 \leq j \leq m\} = k + \lfloor (n-k)/k \rfloor + 1.$$

We then obtain the desired result by applying Corollary 5.12:

$$\lambda(G) \leq k + \lfloor (n-k)/k \rfloor + \frac{n - \sum_{i=1}^{r-1} \hat{y}_i}{\hat{y}_r} = k + \frac{n-k}{k}.$$

∎

We will now prove that systematic MDS matrices achieve the bound of Corollary 6.6 with equality in the case $n \geq 2k$. We start by introducing some objects that we will need to prove the result.

Notation 6.7. *Let $a, b \in \mathbb{Z}$ such that $2 \leq a < b$. Assume $q > b$ and let α be a primitive element of $\mathbb{F}_q$. Define the matrix*

$$G_\alpha^{a,b} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \alpha & \alpha^2 & \cdots & \alpha^{b-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha^{(a-1)} & \alpha^{2(a-1)} & \cdots & \alpha^{(b-1)(a-1)} \end{pmatrix} \in \mathbb{F}_q^{a\times b}.$$

Note that having $q > b$ is necessary and sufficient for the columns of $G_\alpha^{a,b}$ to be pairwise distinct. Sufficiency can be seen by considering the second row of $G_\alpha^{a,b}$ and the fact that the multiplicative order of α is $q-1$. Necessity follows from the fact that if $q \le b$ then at least one of the columns of $G_\alpha^{a,b}$ indexed by $\{2,\dots,b\}$ is equal to the first column. Moreover, the matrix $G_\alpha^{a,b}$ is a generator matrix of a *Reed–Solomon code* [24], which is a type of MDS code; see [21].

Lemma 6.8. *Following Notation 6.7, let $\mathcal{R} = \mathcal{R}^{\min}(G_\alpha^{a,b})$. The following hold.*

1. *Let $i \in \{1,\dots,a\}$ and $R \subseteq \{1,\dots,b\}$. Then $R \in \mathcal{R}_i$ if and only if $|R| = a$.*
2. *Let $\nu \in \{1,\dots,b\}$. We have*

$$|\{R \in \mathcal{R}_i \mid i \in \{1,\dots,a\}, \nu \in R\}| = \binom{b-1}{a-1}.$$

Proof. We first observe that the second part of the lemma follows from the first and the fact that

$$\{S \subseteq \{1,\dots,b\} \mid |S| = a,\ \nu \in S\} = \binom{b-1}{a-1} \text{ for all } 1 \le \nu \le b.$$

To prove the first part, let $G = G_\alpha^{a,b}$. Suppose that $R \in \mathcal{R}_i$ and let us prove $|R| = a$. We first show that $|R| \le a$. Assume towards a contradiction that $|R| > a$. By Proposition B.3, there exists $R' \subseteq R$ such that $|R| = a$ and $R' \in \mathcal{R}_i$, which contradicts the fact that R is i-minimal. We now show that $|R| \ge a$. Towards a contradiction, assume $|R| = c < a$. Because of the structure of G, we can assume without loss of generality $i = a$ and $R = \{1,\dots,c\}$. We will prove that $R \notin \mathcal{R}_a$, which is a contradiction. That is equivalent to showing that $e_a \notin \langle G^\nu \mid \nu \in \{1,\dots,c\}\rangle$, which can be seen from the fact that the matrices

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \alpha & \alpha^2 & \cdots & \alpha^{c-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha^{(a-1)} & \alpha^{2(a-1)} & \cdots & \alpha^{(c-1)(a-1)} \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 & 0 \\ 1 & \alpha & \alpha^2 & \cdots & \alpha^{c-1} & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \alpha^{(a-1)} & \alpha^{2(a-1)} & \cdots & \alpha^{(c-1)(a-1)} & 1 \end{pmatrix}$$

have different ranks. For the other direction, assume $|R| = a$. Since G is an MDS matrix, we have $R \in \mathcal{R}_i^{\text{all}}(G)$ by Proposition B.3. To see that $R \in \mathcal{R}_i$ it is enough to show that all elements of $\mathcal{R}_i$ have cardinality a, which we proved already in the first part of the proof. ∎

Note that the second property of the previous lemma states that each column index of $G_\alpha^{a,b}$ participates in $\binom{b-1}{a-1}$ recovery sets of the system $\mathcal{R}(G_\alpha^{a,b})$.

Theorem 6.9. *Let $(\lambda_1,\dots,\lambda_a) \in \mathbb{R}_{\ge 0}^a$. Following Notation 6.7, if $\lambda_1 + \cdots + \lambda_a \le b/a$, then $(\lambda_1,\dots,\lambda_a) \in \Lambda(G_\alpha^{a,b})$.*

Proof. Let $\mathcal{R} = \mathcal{R}^{\min}(G_\alpha^{a,b})$. For all $i \in \{1,\dots,a\}$ and $R \in \mathcal{R}_i$, let

$$\lambda_{i,R} = \frac{\lambda_i}{\binom{b}{a}} = \lambda_i \frac{a}{b} \frac{1}{\binom{b-1}{a-1}}.$$

We now show that the constraints in Definition 1.4 hold. Constraint (1.3) holds by definition. Constraint (1.1) is satisfied because

$$\sum_{R\in\mathcal{R}_i}\lambda_{i,R}=\lambda_i\frac{a}{b}\frac{1}{\binom{b-1}{a-1}}|\mathcal{R}_i|=\lambda_i\frac{a}{b}\frac{1}{\binom{b-1}{a-1}}\binom{b}{a}=\lambda_i$$

for all $i\in\{1,\dots,a\}$, where the fact that $|\mathcal{R}_i|=\binom{b}{a}$ follows from the first part of Lemma 6.8. By the second part of Lemma 6.8, constraint (1.2) reads as

$$\binom{b-1}{a-1}\sum_{i=1}^{a}\lambda_i\frac{a}{b}\frac{1}{\binom{b-1}{a-1}}\le 1, \tag{6.3}$$

which holds by the theorem's assumption $\lambda_1+\cdots+\lambda_a\le b/a$, since

$$\binom{b-1}{a-1}\sum_{i=1}^{a}\lambda_i\frac{a}{b}\frac{1}{\binom{b-1}{a-1}}=\frac{a}{b}(\lambda_1+\cdots+\lambda_a)\le\frac{a}{b}\frac{b}{a}=1.$$

■

We can now show that systematic MDS matrices with $n\ge 2k$ achieve the bound of Corollary 6.6 (cf. [2]).

Theorem 6.10. *Suppose $n\ge 2k$. If $G\in\mathbb{F}_q^{k\times n}$ is a systematic MDS matrix, then $\lambda(G)=k+\frac{n-k}{k}$.*

Proof. All systematic MDS matrices with $n\ge 2k$ have the same service rate region; see Remark 5.4. Therefore it suffices to prove the result for $G=[\mathrm{Id}_k\mid G_\alpha^{k,n-k}]\in\mathbb{F}_q^{k\times n}$, where Id_k is the $k\times k$ identity matrix over $\mathbb{F}_q$ and $G_\alpha^{k,n-k}$ is as in Notation 6.7.

The fact $\lambda_1+\cdots+\lambda_k\le k+(n-k)/k$ follows from Corollary 6.6. Let $\lambda_{i,\{i\}}=1$ for all $i\in\{1,\dots,k\}$, and observe that $((n-k)/k,0,\dots,0)\in\Lambda(G_\alpha^{k,n-k})$ by taking $a=k$ and $b=n-k$ in Theorem 6.9. Then $((n-k)/k+1,1,\dots,1)\in\Lambda(G)$ by the definition of G. ■

We conclude this section by noting that Corollary 6.6 is not necessarily met with equality when $n<2k$, or if G is not systematic. For the case where $n<2k$, see, for instance, Example 5.13. For the case where $n\ge 2k$ and G is a nonsystematic MDS matrix, consider

$$G=\begin{pmatrix}2&1&3&4\\1&2&3&5\end{pmatrix}\in\mathbb{F}_7^{2\times 4}.$$

Then $\lambda(G)=2<3$.

Appendix A. Polytopes. In this appendix, we collect some background material about polytopes and their properties. More details can be found in standard references; see, e.g., [19, 26, 32].

We start by recalling that a *polyhedron* is a set of the form $\mathcal{P}=\{x\in\mathbb{R}^m\mid Ax^\top\le b^\top\}$, where $A\in\mathbb{R}^{\ell\times m}$, $\ell,m\ge 1$, $b\in\mathbb{R}^\ell$, and $\le$ is applied componentwise. Such a set $\mathcal{P}$ is called a *polytope* if it is bounded. A fundamental result on polyhedra states that every polytope is the convex hull of a finite set of points. For a (possibly infinite) set $S\subseteq\mathbb{R}^m$, we let $\mathrm{conv}(S)$ denote its convex hull, where $\mathrm{conv}(\emptyset)=\emptyset$.

Theorem A.1 (see e.g. [14]). *Let $\mathcal{P} \subseteq \mathbb{R}^m$ be a polytope. Then a finite set $S \subseteq \mathbb{R}^m$ exists, such as $\mathcal{P} = \operatorname{conv}(S)$.*

A *vertex* of a polytope $\mathcal{P} \subseteq \mathbb{R}^m$ is an element $v \in \mathcal{P}$ with $v \notin \operatorname{conv}(\mathcal{P} \setminus \{v\})$. The set of vertices of $\mathcal{P}$ is denoted by $|(\mathcal{P})$. Note that if $\mathcal{P} = \operatorname{conv}(S)$ is a poltyope, then $|(\mathcal{P}) \subseteq S$. Thus $\mathcal{P} = \operatorname{conv}(|(\mathcal{P}))$. Moreover, a nonempty polytope has at least one vertex.

We recall the following crucial property of vertices.

Proposition A.2. *Let $\mathcal{P} = \{x \in \mathbb{R}^m \mid Ax^\top \leq b^\top\}$ be a polytope, with $A \in \mathbb{R}^{\ell \times m}$. Let v be a vertex of $\mathcal{P}$. Then there exists $I \subseteq \{1, \dots, \ell\}$ such that $\operatorname{rank}(A[I]) = m$ and $\{v\} = \{x \in \mathbb{R}^m \mid A[I]x^\top = b[I]^\top\}$, where $A[I]$ and $b[I]$ are obtained from A and b by deleting the rows and components (respectively) not indexed by I.*

The previous result remarkably shows that rational polytopes have *rational* vertices (i.e., with rational entries). Recall that a polyhedron of the form $\{x \in \mathbb{R}^m \mid Ax \leq b^\top\}$ with $A \in \mathbb{Q}^{\ell \times m}$ and $b \in \mathbb{Q}^\ell$ is called *rational*. Then Proposition A.2 combined with Gaussian elimination leads to the following result.

Corollary A.3. *A rational polytope has rational vertices.*

We conclude this appendix by recalling that a polytope $\mathcal{P} \subseteq \mathbb{R}^m$ is *down-monotone* if $x \geq 0$ for all $x \in \mathcal{P}$ and for all $y \in \mathbb{R}^m$ and $x \in \mathcal{P}$ with $0 \leq y \leq x$ we have $y \in \mathcal{P}$. All polytopes in this paper are down-monotone.

Appendix B. Error-correcting codes.

Definition B.1. *An $[n,k]_q$ (*error-correcting*) code is a k-dimensional $\mathbb{F}_q$-linear subspace $\mathcal{C} \leq \mathbb{F}_q^n$. We call n the* length *of $\mathcal{C}$. A matrix $G \in \mathbb{F}_q^{k \times n}$ whose rows span $\mathcal{C}$ is called a* generator matrix *for $\mathcal{C}$. The $[n, n-k]_q$ code $\mathcal{C}^\perp = \{x \in \mathbb{F}_q^n \mid xy^\top = 0 \text{ for all } y \in \mathcal{C}\} \leq \mathbb{F}_q^n$ is the* dual *of $\mathcal{C}$.*

The error correction capability of $\mathcal{C}$ is measured by a fundamental parameter defined as follows.

Definition B.2. *The* Hamming weight *of a vector $x \in \mathbb{F}_q^n$ is the integer $\operatorname{wt}^{\mathrm{H}}(x) = |\{i \mid x_i \neq 0\}|$. The* minimum (Hamming) distance *of a code $\mathcal{C} \leq \mathbb{F}_q^n$ is $d^{\mathrm{H}}(\mathcal{C}) = \min\{\operatorname{wt}^{\mathrm{H}}(x) \mid x \in \mathcal{C}, x \neq 0\}$.*

This paper mainly focuses on $[n,k]_q$ codes with $k + d - 1 = n$. Such codes are called *MDS* (maximum distance separable). A full rank matrix that generates an MDS code is called an *MDS matrix*. These matrices are known to exist only over sufficiently large finite fields ($q \geq n - 1$ suffices). Determining for which field sizes MDS matrices exist has been an open problem since 1955; see [27]. We conclude with the following handy characterization of MDS matrices. The proof can be found in [21, page 318].

Proposition B.3. *Let $G \in \mathbb{F}_q^{k \times n}$ be a matrix. Then G is an MDS matrix if and only if every k columns of G are $\mathbb{F}_q$-linearly independent.*

Acknowledgments. The authors would like to thank Laura Sanità for fruitful discussions on combinatorial optimization and down-monotone polytopes, and the anonymous referees for their comments and suggestions.

REFERENCES

[1] M. Aktaş, S. E. Anderson, A. Johnston, G. Joshi, S. Kadhe, G. L. Matthews, C. Mayer, and E. Soljanin, *On the service capacity region of accessing erasure coded content*, in 55th Annual Allerton Conference on Communication, Control, and Computing, 2017.

[2] M. Aktas, G. Joshi, S. Kadhe, F. Kazemi, and E. Soljanin, *Service rate region: A new aspect of coded distributed system design*, IEEE Trans. Inform. Theory, 67 (2022), pp. 7940–7963.

[3] G. N. Alfarano, A. Ravagnani, and E. Soljanin, *Dual-code bounds on multiple concurrent (local) data recovery*, in IEEE International Symposium on Information Theory, 2022.

[4] S. E. Anderson, A. Johnston, G. Joshi, G. L. Matthews, C. Mayer, and E. Soljanin, *Service capacity region of content access from erasure coded storage*, in IEEE Information Theory Workshop, 2018.

[5] T. M. Apostol, *Calculus*, *Vol.* 1, 2nd ed., John Wiley & Sons, 1991.

[6] D. L. Barrow and P. W. Smith, *Spline notation applied to a volume problem*, Amer. Math. Monthly, 86 (1979), pp. 50–51.

[7] R. Bhatia and C. Davis, *A better bound on the variance*, Amer. Math. Monthly, 107 (2000), pp. 353–357.

[8] J. Cohen and T. Hickey, *Two algorithms for determining volumes of convex polyhedra*, J. ACM, 26 (1979), pp. 401–414.

[9] G. B. Dantzig, *Discrete-variable extremum problems*, Oper. Res., 5 (1957), pp. 266–288.

[10] A. G. Dimakis, K. Ramchandran, Y. Wu, and C. Suh, *A survey on network codes for distributed storage*, Proc. IEEE, 99 (2011), pp. 476–489.

[11] A. Fazeli, A. Vardy, and E. Yaakobi, *Codes for distributed PIR with low storage overhead*, in IEEE International Symposium on Information Theory, 2015, pp. 2852–2856.

[12] A. Fazeli, A. Vardy, and E. Yaakobi, *Codes for distributed PIR with low storage overhead*, in 2015 IEEE International Symposium on Information Theory, IEEE, 2015.

[13] Z. Füredi and I. Bárány, *Computing the volume is difficult*, in Proceedings of the 18th Annual ACM Symposium on Theory of Computing, 1986.

[14] B. Grünbaum, V. Klee, M. A. Perles, and G. C. Shephard, *Convex Polytopes*, Pure Appl. Math. 16, Springer, 1967.

[15] Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai, *Batch codes and their applications*, in Proceedings of the 36th Annual ACM Symposium on Theory of Computing, 2004.

[16] F. Kazemi, E. Karimi, E. Soljanin, and A. Sprintson, *A combinatorial view of the service rates of codes problem, its equivalence to fractional matching and its connection with batch codes*, in IEEE International Symposium on Information Theory, 2020.

[17] F. Kazemi, S. Kurz, and E. Soljanin, *A geometric view of the service rates of codes problem and its application to the service rate of the first order Reed-Muller code*, in IEEE International Symposium on Information Theory, 2020.

[18] F. Kazemi, S. Kurz, and E. Soljanin, *Efficient storage schemes for desired service rate regions*, in IEEE Information Theory Workshop, 2021.

[19] B. Korte and J. Vygen, *Combinatorial Optimization: Theory and Algorithms*, 6th ed., Algorithms and Combinatorics 21, Springer, 2018.

[20] R. Lang, *A note on the measurability of convex sets*, Arch. Math., 47 (1986), pp. 90–92.

[21] J. MacWilliams and N. Sloane, *The Theory of Error-Correcting Codes*, Elsevier, 1977.

[22] J. R. Munkres, *Topology*, 2nd ed., Prentice Hall, 2000.

[23] D. A. Patterson, G. Gibson, and R. H. Katz, *A case for redundant arrays of inexpensive disks (RAID)*, in Proceedings of the ACM SIGMOD, 1988.

[24] I. S. Reed and G. Solomon, *Polynomial codes over certain finite fields*, J. Soc. Indust. Appl. Math., 8 (1960), pp. 300–304, https://doi.org/10.1137/0108018.

[25] A.-E. Riet, V. Skachek, and E. K. Thomas, *Asynchronous batch and PIR codes from hypergraphs*, in IEEE Information Theory Workshop, 2018.

[26] A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley & Sons, 1998.

[27] B. Segre, *Curve razionali normali e k-archi negli spazi finiti*, Ann. Mat. Pura Appl., 39 (1955), pp. 357–379.

[28] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, *The Hadoop distributed file system*, in IEEE Symposium on Mass Storage Systems and Technologies, 2010.
[29] V. Skachek, *Batch and PIR codes and their connections to locally repairable codes*, in Network Coding and Subspace Designs, Springer, 2018, pp. 427–442.
[30] I. Tamo and A. Barg, *A family of optimal locally recoverable codes*, IEEE Trans. Inform. Theory, 60 (2014), pp. 4661–4676.
[31] M. A. Tsfasman and S. G. Vlăduţ, *Algebraic-Geometric Codes*, Kluwer Academic, 1991.
[32] G. M. Ziegler, *Lectures on Polytopes*, Grad. Texts in Math. 152, Springer, 2012.