Optimal rounding under integer constraints

Rama Cont* and Massoud Heidari†

2026

Abstract

Given N real numbers whose sum is an integer, we study the problem of finding N integers that preserve the sum while minimizing the rounding error. We first show that every optimal solution necessarily rounds each coordinate either to its floor or its ceiling, reducing the problem to the selection of the coordinates to be rounded upward. This characterization extends to a class of separable convex integer optimization problems with a single sum constraint.

For the resulting optimization problem we characterize the complete set of optimal solutions and show that rounding upward the largest fractional parts simultaneously minimizes the L^q rounding error for every $q \geq 1$. More generally, the resulting error vector is minimal in the weak-majorization order and therefore minimizes every symmetric convex coordinatewise nondecreasing loss of the rounding error. When the L^q-optimal solution is not unique, we provide an explicit tie-breaking rule that minimizes the relative rounding error among all optimal solutions.

These structural results lead to a deterministic algorithm with linear $O(N)$ worst-case complexity. Unlike independent randomized rounding, which preserves the target coordinates and the sum constraint only in expectation, the proposed method computes an exactly feasible, provably optimal integer rounding with deterministic optimality guarantee. Besides solving the constrained rounding problem, the algorithm applies as the rounding step in relaxed integer optimization problems with a single conservation constraint.

Keywords: integer optimization, optimal rounding, randomized rounding, rounding heuristic, mixed integer programming, apportionment problem, quota methods, resource allocation.

Mathematics Subject Classification: 90C10 90C11 97N20

*Mathematical Institute, University of Oxford. E-mail: Rama.Cont@maths.ox.ac.uk

†LeastBit Information Lab.

Contents

1 Introduction **3**
1.1 Related literature 3
1.2 Outline 4

2 Optimal rounding under integer constraints **5**
2.1 Notations 5
2.2 Problem set-up 5
2.3 Related problems and ramifications 6
2.4 Applications and scope 8

3 Structure and optimality of solutions **9**
3.1 Reduction to componentwise rounding 9
3.2 Characterization by largest fractional parts 10
3.3 Relative error tie-breaking 12
3.4 Extension to common-loss separable objectives 13
3.5 Weak majorization and universal optimality 14

4 Algorithms for optimal rounding under integer constraints **21**
4.1 Implementation via sorting 21
4.2 Deterministic implementation with linear complexity 23

5 Comparison with other rounding methods **27**
5.1 Fractional rounding 27
5.2 Independent randomized rounding 28
5.3 Dependent randomized rounding 30

6 Matroid-constrained optimal rounding **31**

1 Introduction

The rounding, or integer approximation, of real numbers is a key step in many algorithms used in integer optimization [10, 15, 17, 26], whereby

(i) an optimization problem over integers is replaced by a corresponding problem over real numbers ('relaxation' step);

(ii) the real solution obtained in (i) is 'rounded' to obtain an integer solution satisfying the constraints of the original problem (rounding step).

If the objective function of the relaxed problem is linear, convex or has some other special property, step (i) may be solved using efficient numerical methods with polynomial complexity in the dimension N of the problem. The rounding step, on the other hand, may be described as a constrained integer optimization problem over $\{0,1\}^N$ for which no generic polynomial algorithm is known.

Given the large dimensionality of integer programming problems arising in many applications, exhaustive search on $\{0,1\}^N$ is certainly not an option. Randomized rounding has been extensively used and analyzed as an alternative [24]. Yet in many instances randomized rounding may result in a substantial loss in accuracy [26]. Thus, in practice, the final rounding step sets a lower bound on the overall accuracy of many integer optimization algorithms, which illustrates the importance of good rounding algorithms.

Many integer optimization problems, such as scheduling of tasks on parallel machines [22, 17, 25], or the proportional allocation of seats to political parties in elections with party list voting systems [1, 19], are naturally subject to integer constraints. The latter problem, known as the *apportionment problem*, has a long history in mathematics, dating back to Polya [21, 20], and continues to generate a lot of interest [9, 14]. Relaxation to real variables followed by rounding each component to the nearest integer may typically fail to satisfy such constraints and many different –and non-equivalent– methods exist for obtaining approximate or exact solutions satisfying the integer constraints [1, 24, 19, 26]. Most of these methods are based on *heuristics* which provide approximate solutions. Some algorithms, based on randomized rounding may yield a bias in finite samples, which may or may not vanish asymptotically [5, 14]. Even in the case where they yield exact solutions, the mathematical properties of such algorithms have not been systematically analyzed until recently [19].

We complement this picture by demonstrating the link between various families of such *optimal rounding* problems, proposing an algorithm with deterministic linear complexity for solving them, giving a systematic analysis of the optimality and computational complexity of this algorithm and comparing it to some popular methods.

1.1 Related literature

The problem considered here is related to several strands of the literature on integer optimization, constrained rounding, and resource allocation.

A common paradigm in mixed-integer optimization is to solve a continuous relaxation and subsequently round the relaxed solution. Berthold [2] introduced the RENS heuristic, which restricts integer variables to their floor or ceiling values around a relaxation point and solves the resulting mixed-integer subproblem. The problem considered here is a structured special case in which the only constraint is preservation of the total sum and the objective is the rounding error. In this setting we show that the floor-or-ceiling restriction is not a heuristic but a *necessary* property of every optimum, allowing the subproblem to be solved explicitly in linear time.

The common-loss separable subclass of Problem 2 considered in Proposition 3.5 belongs to the class of separable convex resource allocation problems.

Polynomial algorithms for such problems have been studied by Federgruen and Groenevelt [6], Ibaraki and Katoh [13], and Hochbaum and Shanthikumar [12]. Our contribution is complementary to these general results. Exploiting the common-loss structure and the equality constraint, we obtain a complete characterization of the optimal integer solutions, prove simultaneous optimality for every L^q norm, establish weak-majorization optimality, and derive a deterministic $O(N)$ algorithm.

The paper also relates to the literature on randomized rounding. Independent randomized rounding [24] preserves the target point only in expectation and does not, in general, satisfy the sum constraint. More sophisticated dependent rounding schemes [8] preserve cardinality or other linear constraints while retaining prescribed marginals. Unlike these randomized methods, our algorithm is deterministic and computes an exact minimizer of the constrained rounding problem.

The decimal-rounding problem considered in Section 2.3 is a single-total instance of controlled rounding, introduced by Cox and Ernst [4], while the preservation of cumulative sums studied by Knuth's two-way rounding algorithm [16] addresses a different rounding criterion based on discrepancy rather than optimal approximation in an L^q norm.

Finally, our primary rounding rule coincides with the classical largest-remainder (Hare–Niemeyer) method used in proportional apportionment [1, 23]. The present work provides an optimization interpretation of this rule by showing that it minimizes every L^q rounding error and, more generally, every symmetric, convex, coordinatewise nondecreasing function of the absolute rounding-error vector. The proposed relative-error tie-breaking rule further selects a relative-error minimizer among multiple L^q-optimal allocations.

1.2 Outline

We formulate the *optimal rounding* problem in Section 2. In Section 3, we establish the componentwise rounding structure (Proposition 3.1), characterize the set of optimal solutions (Proposition 3.2), derive a relative-error tie-breaking rule (Corollary 3.4), extend the characterization to common-loss separable objectives, and prove weak-majorization optimality (Theorem 3.9). Section 4 presents sorting-based and deterministic linear-time implementations and analyzes their

computational complexity. Finally, Section 5 compares the proposed method with commonly used deterministic and randomized rounding procedures.

2 Optimal rounding under integer constraints

2.1 Notations

Denote by $\mathbb{Z}$ the set of integers, $\mathbb{N}$ the set of non-negative integers and $\mathbb{R}$ the set of real numbers. For a vector $x = (x_1, ..., x_N) \in \mathbb{R}^N$ and $q \geq 1$, denote

$$\|x\|_q = \left(\sum_{i=1}^{N} |x_i|^q \right)^{1/q}$$

For a real number $y \in \mathbb{R}$, we denote

$$\lfloor y \rfloor = \sup\{m \in \mathbb{Z}, \quad m \leq y\} \quad \lceil y \rceil = \inf\{m \in \mathbb{Z}, \quad m \geq y\}$$

For $(z_1, z_2, .., z_N) \in \mathbb{R}^N$ we denote the *ordered* sequence as

$$z_{(1)} \geq \cdots \geq z_{(N)}.$$

2.2 Problem set-up

The problem of optimal rounding under integer constraints can be formulated in the following way:

Problem 1 (Optimal rounding under integer constraints (**ORIC**))**.**
Let $q \geq 1$. Given positive real numbers $x = (x_1, ..., x_N) \in \mathbb{R}_+^N$ with

$$\sum_{i=1}^{N} x_i = M \in \mathbb{N} \tag{1}$$

find a set of integers $m = (m_1, ..., m_N) \in \mathbb{N}^N$ which minimizes

$$\|x - m\|_q = \left(\sum_{i=1}^{N} |x_i - m_i|^q \right)^{1/q} \quad \text{under} \quad \sum_{i=1}^{N} m_i = M. \tag{2}$$

We denote the corresponding (L^q) rounding error

$$V_q(x) = \inf\{\|x - m\|_q, \qquad m \in \mathbb{N}^N, \sum_{i=1}^{N} m_i = M\}. \tag{3}$$

The non-trivial feature of the problem is the presence of the integer constraint. As the example $x = (2.25, 3.4, 4.35)$ shows, componentwise rounding to the nearest integer may fail to satisfy such a constraint.

Problem 1 is a special case of the following integer programming problem.

Problem 2. *Given a convex function $f : \mathbb{R}^N \to \mathbb{R}$ find a set of integers $m = (m_1, ..., m_N) \in \mathbb{N}^N$ which minimizes*

$$\inf_{m \in \mathbb{N}^N} f(m) \quad \text{under} \qquad \sum_{i=1}^{N} m_i = M \in \mathbb{N}. \tag{4}$$

This is an optimization problem over the finite set

$$\{m \in \mathbb{N}^N, \sum_{i=1}^{N} m_i = M\}$$

so the infimum is always attained. We denote by

$$V(f) = \min\{f(m), \qquad m \in \mathbb{N}^N, \sum_{i=1}^{N} m_i = M\} \tag{5}$$

the value of this minimum. The feasible set has cardinality

$$\binom{M+N-1}{N-1},$$

which can grow exponentially with N when M grows proportionally with N. So, at first sight, Problems 1 and 2 appear to be integer optimization problems with exponential complexity.

Our contribution is to study the structure of these problems and show how some of them can be solved using an algorithm with polynomial -in fact *linear*- complexity in N. We propose such a solution for Problem 1 and the common-loss separable subclass of Problem 2, described in Proposition 3.5.

We first show in Section 3 that, notwithstanding the constraint, the solution necessarily consists in rounding each component either up or down (Proposition 3.1). We can thus reformulate the problem as an optimization problem on $\{0, 1\}^N$. Next, in Section 4 we first describe a sorting-based implementation with $O(N \log N)$ complexity and then give a deterministic linear-time implementation with $O(N)$ worst-case complexity.

2.3 Related problems and ramifications

Problem 1 is a 'pure integer programming' problem in the sense that the relaxation to the case where $m \in \mathbb{R}^N$ is trivial. As such, it enters as a building block in many integer and mixed-integer programming problems in which one first solves a relaxation of the problem to real variables then projects back the solution of the relaxed problem onto $\mathbb{Z}^N$.

The following problem arises e.g. in rounding problems encountered in accounting, where one rounds N entries while leaving the total unchanged up to the nearest dollar:

Problem 3 (Decimal rounding with exact preservation of total)**.** *Let $q \geq 1$. Given $x = (x_1, \ldots, x_N) \in \mathbb{R}_+^N$ and $k \in \mathbb{N}$ such that*

$$10^k \sum_{i=1}^{N} x_i = M \in \mathbb{N},$$

find $y = (y_1, \ldots, y_N) \in 10^{-k}\mathbb{N}^N$ which minimizes

$$\sum_{i=1}^{N} |x_i - y_i|^q \quad \text{subject to} \qquad \sum_{i=1}^{N} y_i = \sum_{i=1}^{N} x_i = \frac{M}{10^k}.$$

This problem is equivalent to Problem 1. Indeed, define

$$X_i = 10^k x_i \qquad \text{and} \qquad m_i = 10^k y_i.$$

Then $m \in \mathbb{N}^N$ and the constraint becomes

$$\sum_{i=1}^{N} m_i = M = \sum_{i=1}^{N} X_i.$$

Moreover,

$$\sum_{i=1}^{N} |x_i - y_i|^q = 10^{-kq} \sum_{i=1}^{N} |X_i - m_i|^q.$$

So minimizing this objective is equivalent to minimizing

$$\sum_{i=1}^{N} |X_i - m_i|^q \quad \text{under} \quad m \in \mathbb{N}^N, \qquad \sum_{i=1}^{N} m_i = M,$$

which is precisely Problem 1 applied to $X = 10^k x$.

Problem 1 has multiple solutions precisely when there is a tie at the rounding cutoff, that is, when

$$1 \leq I \leq N - 1 \qquad \text{and} \qquad z_{(I)} = z_{(I+1)},$$

where $z_{(1)} \geq \cdots \geq z_{(N)}$ are the fractional parts in decreasing order. In this case one might consider minimizing the relative rounding error among all solutions of Problem 1:

Problem 4 (Optimal rounding with smallest relative error)**.** *Let $x_i > 0$ and*

$$\mathcal{M}_q(x) = \operatorname*{arg\,min}_{\substack{m \in \mathbb{N}^N \\ \sum_{i=1}^N m_i = M}} \sum_{i=1}^{N} |x_i - m_i|^q.$$

Find $m^* \in \mathcal{M}_q(x)$ such that

$$\sum_{i=1}^{N} \frac{|x_i - m_i^*|^q}{x_i^q} = \min_{m \in \mathcal{M}_q(x)} \sum_{i=1}^{N} \frac{|x_i - m_i|^q}{x_i^q}.$$

2.4 Applications and scope

Problem 1 arises whenever M identical, indivisible units must be distributed among N recipients according to a real-valued target allocation

$$x = (x_1, \ldots, x_N), \qquad x_i \geq 0, \qquad \sum_{i=1}^{N} x_i = M.$$

The objective is to construct an integer allocation m that preserves the total number of units while remaining as close as possible to the target allocation.

Example 2.1 (Apportionment [23])**.** Let $v_i > 0$ denote the number of votes received by party i, or the population associated with constituency i, and let M be the number of seats to be allocated. The exact quota is

$$x_i = M \frac{v_i}{\sum_{j=1}^{N} v_j}, \qquad \sum_{i=1}^{N} x_i = M.$$

Problem 1 asks for an integer seat allocation m satisfying

$$\sum_{i=1}^{N} m_i = M$$

and minimizing the discrepancy from the exact quota vector. The ORIC allocation detailed in Section 3 coincides with the largest-remainder, or Hare–Niemeyer [19], apportionment. The optimization properties established below concern closeness to the quota vector; other axiomatic properties of apportionment methods constitute separate criteria.

The same mathematical problem arises when election seats are to be apportioned across constituencies according to their populations. See Balinski and Young [1] for a detailed exposition and Grimmett [9] for a discussion in the context of the European Parliament. This problem has a long history in mathematics, dating back to Polya [21, 20], and continues to generate a lot of interest [19, 9, 14].

Example 2.2 (Allocation of identical units)**.** *Suppose that M identical production, storage, computational, or investment units are to be allocated among N activities. If x_i denotes the desired real-valued allocation to activity i, Problem 1 gives an integer allocation preserving the total number of units and minimizing the rounding error.*

The method also applies as a rounding subroutine when the residual integer feasibility condition of a relaxed optimization problem reduces to a single conservation or cardinality constraint. General portfolio, assignment, and scheduling problems typically involve additional weighted, capacity, or coupling constraints and therefore do not reduce directly to Problem 1.

3 Structure and optimality of solutions

We first establish that every solution to Problem 1 rounds each coordinate either to its floor or to its ceiling. We then characterize the complete set of optimal solutions through the largest fractional parts and derive a relative-error tie-breaking rule when the optimum is not unique. Next, we extend the largest-fractional-parts characterization to the common-loss separable subclass of Problem 2, considered in Proposition 3.5. Finally, we show that the absolute-error vector of an ORIC solution is minimal in the weak-majorization order.

3.1 Reduction to componentwise rounding

Proposition 3.1 (Reduction to term-by-term rounding)**.**
Let $x \notin \mathbb{N}^N$ and $m^* = (m_1^*, ..., m_N^*) \in \mathbb{N}^N$ be a solution to Problem 1. Then

$$\forall i = 1..N, \qquad m_i^* \in \{\lfloor x_i \rfloor, \lceil x_i \rceil\}. \tag{6}$$

Proof. Define the feasible set $\mathcal{F} = \left\{ m \in \mathbb{N}^N : \sum_{i=1}^N m_i = M \right\}$ and

$$F_q(m) = \sum_{i=1}^{N} |m_i - x_i|^q.$$

As the feasible set $\mathcal{F}$ is non-empty and finite, F_q attains its minimum and the set of minimizers is non-empty. We now show that every minimizer m rounds each coordinate either downward or upward. Suppose that a minimizer m has a coordinate satisfying $m_i < \lfloor x_i \rfloor$. Because

$$\sum_{\ell=1}^{N} (m_\ell - x_\ell) = 0,$$

there exists an index j such that $m_j > x_j$. Set

$$A = x_i - m_i \geq 1, \qquad B = m_j - x_j > 0,$$

and define $m' = m + e_i - e_j$, where e_i denotes the ith standard basis vector. Then $m' \in \mathcal{F}$. Moreover,

$$F_q(m) - F_q(m') = A^q - (A-1)^q + B^q - |B-1|^q.$$

Since $A \geq 1$ and $q \geq 1$,

$$A^q - (A-1)^q \geq 1.$$

If $0 < B < 1$, then

$$B^q - |B-1|^q = B^q - (1-B)^q > -1,$$

whereas if $B \geq 1$, then

$$B^q - |B-1|^q \geq 0.$$

Hence $F_q(m) - F_q(m') > 0$, contradicting the optimality of m. Therefore no minimizer can satisfy $m_i < \lfloor x_i \rfloor$. A symmetric argument shows that no minimizer can satisfy $m_i > \lceil x_i \rceil$. □

Proposition 3.1 therefore reduces Problem 1 to a binary optimization problem subject to the cardinality constraint that exactly

$$I = M - \sum_{i=1}^{N} \lfloor x_i \rfloor$$

noninteger components are rounded upward.

3.2 Characterization by largest fractional parts

Define $z_i = x_i - \lfloor x_i \rfloor \in [0, 1)$, and

$$I = M - \sum_{i=1}^{N} \lfloor x_i \rfloor = \sum_{i=1}^{N} z_i \in \mathbb{N}.$$

Since $z_i < 1$ for every i, we have $0 \le I \le N - 1$. By Proposition 3.1, every solution to Problem 1 rounds each component either downward or upward. Hence every solution has the form

$$m_i = \lfloor x_i \rfloor + 1_{\{i \in S\}}, \qquad i = 1, \ldots, N,$$

for some $S \subseteq \{i : z_i > 0\}$. The sum constraint then requires $|S| = I$. Thus, the constrained rounding problem reduces to selecting the I coordinates to be rounded upward. The following result characterizes all solutions of Problem 1.

Proposition 3.2 (Characterization by largest fractional parts)**.**
Let $q \ge 1$, and let

$$x = (x_1, \ldots, x_N) \in \mathbb{R}_+^N, \qquad \sum_{i=1}^{N} x_i = M \in \mathbb{N}.$$

For each i, write $a_i = \lfloor x_i \rfloor$ and $z_i = x_i - a_i$, and let

$$I = M - \sum_{i=1}^{N} a_i = \sum_{i=1}^{N} z_i.$$

Let $S^* \subseteq \{1, \ldots, N\}$ consist of I indices having the largest fractional parts z_i, with ties at the cutoff resolved arbitrarily, and define

$$m_i^* = a_i + 1_{\{i \in S^*\}}, \qquad i = 1, \ldots, N.$$

Then:

1. *$m^* \in \mathbb{N}^N$ and*

$$\sum_{i=1}^{N} m_i^* = M.$$

2. *The vector m^* solves Problem 1:*

$$\sum_{i=1}^{N} |x_i - m_i^*|^q = \min_{\substack{m \in \mathbb{N}^N \\ \sum_{i=1}^N m_i = M}} \sum_{i=1}^{N} |x_i - m_i|^q.$$

   The same choice of S^* is optimal for every $q \geq 1$.

3. *The set of solutions to Problem 1 consists precisely of the vectors obtained by rounding upward I coordinates whose fractional parts are among the I largest fractional parts.*

4. *The solution is unique if and only if either $I = 0$, or $z_{(I)} > z_{(I+1)}$, where $z_{(1)} \geq \cdots \geq z_{(N)}$ are the fractional parts in decreasing order.*

Proof. The feasibility statement follows directly from the definition of m^*:

$$\sum_{i=1}^{N} m_i^* = \sum_{i=1}^{N} a_i + |S^*| = \sum_{i=1}^{N} a_i + I = M.$$

By Proposition 3.1, it is sufficient to consider vectors of the form

$$m_i = a_i + 1_{\{i \in S\}}, \qquad |S| = I.$$

For any such set S,

$$\begin{aligned} \sum_{i=1}^{N} |x_i - m_i|^q &= \sum_{i \notin S} z_i^q + \sum_{i \in S} (1 - z_i)^q \\ &= \sum_{i=1}^{N} z_i^q + \sum_{i \in S} \big((1 - z_i)^q - z_i^q \big). \end{aligned}$$

Define the incremental cost of rounding a coordinate upward by

$$d_q(z) = (1 - z)^q - z^q, \qquad 0 \leq z \leq 1.$$

For $q = 1$,

$$d_1(z) = 1 - 2z,$$

which is strictly decreasing. For $q > 1$,

$$d_q'(z) = -q(1 - z)^{q-1} - q z^{q-1} < 0, \qquad 0 < z < 1.$$

Thus d_q is strictly decreasing on $[0, 1]$ for every $q \geq 1$. Minimizing the objective over all sets S of cardinality I is therefore equivalent to selecting the I smallest values of $d_q(z_i)$, which are precisely the I largest values of z_i. This proves the optimality of m^*. Since the ordering induced by d_q is the ordering of the fractional parts for every $q \geq 1$, the same upward-rounded set is simultaneously optimal for all such q.

The same argument also characterizes all solutions. Every coordinate with fractional part strictly above the cutoff must be rounded upward, every coordinate with fractional part strictly below the cutoff must be rounded downward, and the required number of coordinates at the cutoff may be chosen arbitrarily. Hence the solution is unique when $I = 0$, or when there is a strict separation $z_{(I)} > z_{(I+1)}$. If $z_{(I)} = z_{(I+1)}$, different choices within the cutoff tie set yield distinct solutions with the same objective value. □

We will call 'ORIC solution' to refer to the set of solutions characterized by the result (3) above:

Definition 3.3 (ORIC solution)**.** *We call* ORIC solution *of Problem 1 any vector $m \in \mathbb{N}^N$ obtained by rounding upward I coordinates of x having the I largest fractional parts, where $I = M - \sum_{i=1}^{N} \lfloor x_i \rfloor$.*

This definition is independent of $q \geq 1$.

3.3 Relative error tie-breaking

Corollary 3.4 (Relative-error tie-breaking)**.** *Under the assumptions of Proposition 3.2, suppose that one wishes to minimize*

$$R_q(m) = \sum_{i=1}^{N} \frac{|x_i - m_i|^q}{x_i^q}.$$

among the solutions of Problem 1. Assume $I \geq 1$, and let $t = z_{(I)}$ be the fractional part at the cutoff. Define

$$T = \{i : z_i = t\}, \qquad k = I - \#\{i : z_i > t\}.$$

Thus exactly k indices in T must be rounded upward. A minimizer of R_q among the solutions of Problem 1 is obtained by choosing

$$\begin{cases} \text{the } k \text{ largest values of } x_i \text{ in } T, & t < \frac{1}{2}, \\ \text{any } k \text{ indices in } T, & t = \frac{1}{2}, \\ \text{the } k \text{ smallest values of } x_i \text{ in } T, & t > \frac{1}{2}. \end{cases}$$

Since all indices in T have the same fractional part, ordering by x_i is equivalent to ordering by $\lfloor x_i \rfloor$.

Proof. Among the solutions of Problem 1, only the choice of indices inside the cutoff tie set T is variable. For every $i \in T$, the change in relative error caused by rounding i upward rather than downward is

$$\Delta_i = \frac{(1-t)^q - t^q}{x_i^q}.$$

Exactly k such increments must be selected.

If $t < \frac{1}{2}$, then $(1-t)^q - t^q > 0$, so the k smallest increments are obtained by choosing the k largest values of x_i. If $t > \frac{1}{2}$, then $(1-t)^q - t^q < 0$, so the smallest, that is, most negative, increments are obtained by choosing the k smallest values of x_i. If $t = \frac{1}{2}$, every increment is zero, so every choice of k tied coordinates has the same relative error. □

3.4 Extension to common-loss separable objectives

We now turn to Problem 2. The following result gives a solution to Problem 2 for 'common-loss, separable objectives' i.e. when f has the form

$$f(y) = C + \sum_{i=1}^{N} \phi(y_i - x_i)$$

on the feasible hyperplane $\sum_{i=1}^{N} y_i = M$.

Proposition 3.5. *Let $x^* = (x_1^*, ..., x_N^*) \in \mathbb{R}_+^N$ such that $\sum_{i=1}^{N} x_i^* = M \in \mathbb{N}$. Let $\phi : \mathbb{R} \mapsto \mathbb{R}$ be strictly convex and define*

$$f(y) = \sum_{i=1}^{N} \phi(y_i - x_i^*).$$

Then x^* is the unique solution of

$$\inf_{y \in \mathbb{R}_+^N} f(y) \quad \text{under} \qquad \sum_{i=1}^{N} y_i = M \tag{7}$$

and every solution $m^* = (m_1^*, ..., m_N^*) \in \mathbb{N}^N$ of

$$\inf_{m \in \mathbb{N}^N} f(m) \quad \text{under} \qquad \sum_{i=1}^{N} m_i = M. \tag{8}$$

satisfies

$$\forall i \in [1, N], \qquad m_i^* \in \{\lfloor x_i^* \rfloor, \lceil x_i^* \rceil\}. \tag{9}$$

Let $z_i = x_i^* - \lfloor x_i^* \rfloor$ and

$$I = \sum_{i=1}^{N} z_i = M - \sum_{i=1}^{N} \lfloor x_i^* \rfloor \in \mathbb{N}.$$

Then an integer minimizer is obtained by rounding upward exactly the I components with the largest fractional parts z_i, with cutoffs resolved arbitrarily.

Proof. For any feasible y let $u_i = y_i - x_i^*$. we have $\sum_i u_i = 0$. Jensen's inequality gives

$$\frac{1}{N}\sum_{i=1}^{N}\phi(u_i) \geq \phi(\frac{1}{N}\sum_{i=1}^{N}u_i) = \phi(0).$$

Strict convexity gives equality only when $u_1 = ... = u_N$. Since $\sum_{i=1}^{N} u_i = 0$ this implies $u = 0$. Hence x^* is the unique minimizer for the constrained optimization problem (7).

Let m be a minimizer in (8) and $d_i = m_i - x_i^*$. Suppose $m_i < \lfloor x_i^* \rfloor$. Then $d_i \leq -1$. Since $\sum d_i = 0$ there exists $j \neq i$ with $d_j > 0$. The unit increment function $h(t) = \phi(t+1) - \phi(t)$ is strictly increasing. Because $d_j - d_i > 1$, $h(d_i) < h(d_j - 1)$. Consequently

$$f(m+e_i-e_j)-f(m) = \phi(d_i+1)-\phi(d_i)+\phi(d_j-1)-\phi(d_j) = h(d_i)-h(d_j-1) < 0$$

which contradicts optimality. The same exchange argument with the direction of the transfer reversed, excludes $m_i > \lceil x_i^* \rceil$. Therefore every coordinate is rounded either up or down.

Rounding coordinate i upward changes the objective by

$$\delta_i = \phi(1 - z_i) - \phi(-z_i) = h(-z_i)$$

Since h is strictly increasing, δ_i is strictly decreasing in z_i. Thus the I smallest incremental costs correspond exactly to the I largest fractional parts. □

Taking $\phi(u) = |u|^q$ for $q > 1$ corresponds to the L^q minimization problem under integer constraints:

$$\inf_{m\in\mathbb{N}^N}\sum_{i=1}^{N}|m_i - x_i^*|^q \quad \text{under} \qquad \sum_{i=1}^{N} m_i = M \in \mathbb{N} \tag{10}$$

3.5 Weak majorization and universal optimality

We now show that the optimality property of the ORIC solution is stronger than simultaneous L^q-optimality. In fact, its vector of absolute rounding errors is minimal in the weak-majorization order [18]. For $v = (v_1, \ldots, v_N) \in \mathbb{R}_+^N$, define $T_0(v) = 0$ and

$$T_k(v) = \sum_{r=1}^{k} v_{(r)}, \qquad k = 1, \ldots, N, \quad \text{where} \quad v_{(1)} \geq v_{(2)} \geq \cdots \geq v_{(N)}.$$

Definition 3.6 (Weak submajorization)**.** *For $u, v \in \mathbb{R}_+^N$, we say that u is weakly submajorized by v, and write $u \preceq_{\mathrm{w}} v$, if*

$$T_k(u) \leq T_k(v), \qquad k = 1, \ldots, N.$$

Thus $u \preceq_{\mathrm{w}} v$ means: for every k, the sum of the k largest components of u is no larger than the sum of the k largest components of v.

For a feasible integer vector m, define its absolute-error vector by

$$e(m) = \big(|x_1 - m_1|, \ldots, |x_N - m_N|\big).$$

We first record two elementary observations.

Lemma 3.7 (Embedding property)**.** *Suppose that $u, v \in \mathbb{R}^r_+$ satisfy*

$$u \preceq_{\mathrm{w}} v.$$

Then, for every $w \in \mathbb{R}^s_+$, $(u, w) \preceq_{\mathrm{w}} (v, w)$, where (u, w) denotes the concatenation of u and w.

Proof. For every $k \in \{0, \ldots, r + s\}$, the sum of the k largest components of (u, w) may be written as

$$T_k(u, w) = \max_{\ell} \left\{T_\ell(u) + T_{k-\ell}(w)\right\},$$

where the maximum is over the integers ℓ satisfying $0 \le \ell \le r, \quad 0 \le k - \ell \le s$. Since $T_\ell(u) \le T_\ell(v)$ for every ℓ, it follows that

$$\begin{aligned} T_k(u, w) &= \max_{\ell} \left\{T_\ell(u) + T_{k-\ell}(w)\right\} \\ &\le \max_{\ell} \left\{T_\ell(v) + T_{k-\ell}(w)\right\} = T_k(v, w). \end{aligned}$$

Hence $(u, w) \preceq_{\mathrm{w}} (v, w)$. □

Lemma 3.8 (Two-coordinate improvements)**.** *The following weak-majorization relations hold.*

1. *If $A \ge 1$ and $B > 0$, then*

$$\big(A - 1, |B - 1|\big) \preceq_{\mathrm{w}} (A, B), \qquad A - 1 + |B - 1| < A + B.$$

2. *If $0 \le b < a < 1$, then*

$$(1 - a, b) \preceq_{\mathrm{w}} (a, 1 - b), \qquad 1 - a + b < a + 1 - b.$$

Proof. For two-dimensional nonnegative vectors, weak submajorization is equivalent to comparing both the maximum component and the sum of the two components. For the first assertion, if $B \ge 1$, then

$$A - 1 \le A, \qquad |B - 1| = B - 1 \le B.$$

If $0 < B < 1$, then

$$|B - 1| = 1 - B < 1 \le A.$$

In either case,

$$\max\{A - 1, |B - 1|\} \le \max\{A, B\}.$$

Moreover, if $0 < B < 1$, then

$$A - 1 + |B - 1| = A - B < A + B,$$

whereas if $B \geq 1$, then

$$A - 1 + |B - 1| = A + B - 2 < A + B.$$

This proves the first assertion.

For the second assertion, $b < a$ implies

$$b < a \qquad \text{and} \qquad 1 - a < 1 - b.$$

Consequently,

$$\max\{1 - a, b\} \leq \max\{a, 1 - b\}.$$

Also,

$$1 - a + b = 1 - (a - b) < 1 + (a - b) = a + 1 - b.$$

This proves the second assertion. □

We may now state the principal majorization result.

Theorem 3.9 (Weak-majorization optimality of ORIC)**.** *Let*

$$x = (x_1, \ldots, x_N) \in \mathbb{R}_+^N, \qquad \sum_{i=1}^N x_i = M \in \mathbb{N},$$

and let

$$\mathcal{F} = \left\{ m \in \mathbb{N}^N : \sum_{i=1}^N m_i = M \right\}.$$

For each i, write $a_i = \lfloor x_i \rfloor$, $\quad z_i = x_i - a_i$, and define

$$I = M - \sum_{i=1}^N a_i = \sum_{i=1}^N z_i.$$

Let m^* be any ORIC solution: $m_i^* = a_i + \mathbf{1}_{\{i \in S^*\}}$, where S^* consists of I indices having the largest fractional parts z_i, with ties at the cutoff resolved arbitrarily. Then, for every $m \in \mathcal{F}$,

$$e(m^*) \preceq_{\mathrm{w}} e(m).$$

Equivalently,

$$\sum_{r=1}^k e_{(r)}(m^*) \leq \sum_{r=1}^k e_{(r)}(m), \qquad k = 1, \ldots, N.$$

Moreover,

$$\sum_{i=1}^{N} |x_i - m_i^*| < \sum_{i=1}^{N} |x_i - m_i|$$

for every feasible m which is not an ORIC solution. Consequently, equality in the weak-majorization inequality for $k = N$ holds if and only if m is an ORIC solution.

Proof. Fix $m \in \mathcal{F}$. We transform m into an ORIC solution through a finite sequence of feasible exchanges. At every exchange, the new error vector is weakly submajorized by the preceding error vector.

If $I = 0$, then

$$\sum_{i=1}^{N} z_i = 0.$$

Since $z_i \geq 0$, it follows that $z_i = 0$ for every i, so $x \in \mathbb{N}^N$ and $m^* = x$. Hence

$$e(m^*) = 0 \preceq_{\mathrm{w}} e(m)$$

for every $m \in \mathcal{F}$. Moreover, if $m \neq m^*$, then

$$\sum_{i=1}^{N} |x_i - m_i| > 0 = \sum_{i=1}^{N} |x_i - m_i^*|.$$

We may therefore assume below that $I \geq 1$. Suppose first that $m_i < \lfloor x_i \rfloor$ for some i. Since

$$\sum_{\ell=1}^{N} (m_\ell - x_\ell) = 0,$$

there exists an index j such that $m_j > x_j$. Let

$$A = x_i - m_i \geq 1, \qquad B = m_j - x_j > 0,$$

and define $m' = m + e_i - e_j$. Since $m_j > x_j \geq 0$ and $m_j \in \mathbb{N}$, we have $m_j \geq 1$. Hence $m_j - 1 \in \mathbb{N}$, and therefore $m' \in \mathbb{N}^N$. We also have

$$\sum_{\ell=1}^{N} m'_\ell = \sum_{\ell=1}^{N} m_\ell = M.$$

Thus $m' \in \mathcal{F}$. The two affected errors change from (A, B) to $(A - 1, |B - 1|)$. Lemma 3.8 gives

$$(A - 1, |B - 1|) \preceq_{\mathrm{w}} (A, B).$$

The remaining $N - 2$ errors are unchanged. By Lemma 3.7, $e(m') \preceq_{\mathrm{w}} e(m)$. In addition,

$$\sum_{\ell=1}^{N} |x_\ell - m'_\ell| < \sum_{\ell=1}^{N} |x_\ell - m_\ell|.$$

Suppose instead that $m_j > \lceil x_j \rceil$ for some j. Since the deviations still sum to zero, there exists an index i such that $m_i < x_i$. Set

$$A = m_j - x_j \geq 1, \qquad B = x_i - m_i > 0,$$

and again define

$$m' = m + e_i - e_j.$$

The two affected errors change from (A, B) to $(A-1, |B-1|)$. The same argument therefore gives

$$e(m') \preceq_{\mathrm{w}} e(m)$$

and a strict reduction of the total absolute error.

We repeat these exchanges whenever a coordinate lies below its floor or above its ceiling. Each exchange preserves feasibility and strictly decreases

$$\sum_{i=1}^{N} |x_i - m_i|.$$

Since the feasible set $\mathcal{F}$ is finite, the procedure must terminate. At termination, the resulting feasible vector $\widetilde{m}$ satisfies

$$\widetilde{m}_i \in \{\lfloor x_i \rfloor, \lceil x_i \rceil\}, \qquad i = 1, \ldots, N,$$

and $e(\widetilde{m}) \preceq_{\mathrm{w}} e(m)$. Therefore $\widetilde{m}_i = a_i + \mathbf{1}_{\{i \in S\}}$ for some set $S \subseteq \{i : z_i > 0\}$. The sum constraint gives

$$M = \sum_{i=1}^{N} a_i + |S|,$$

so $|S| = I$. Now suppose that S does not consist of I indices having the largest fractional parts. Then there exist indices

$$i \notin S, \qquad j \in S,$$

such that $z_i > z_j$. The coordinate i is currently rounded downward and has error z_i, while coordinate j is currently rounded upward and has error $1 - z_j$.

Define a new upward-rounded set by

$$S' = (S \setminus \{j\}) \cup \{i\}.$$

The corresponding feasible vector $\widetilde{m}'$ is obtained by rounding i upward and j downward. The affected pair of errors changes from $(z_i, 1 - z_j)$ to $(1 - z_i, z_j)$. Since

$$0 \leq z_j < z_i < 1,$$

Lemma 3.8 gives

$$(1 - z_i, z_j) \preceq_{\mathrm{w}} (z_i, 1 - z_j).$$

After adjoining the unchanged errors, we obtain $e(\widetilde{m}') \preceq_{\mathrm{w}} e(\widetilde{m})$. Furthermore,

$$\begin{aligned}(1 - z_i) + z_j &= 1 - (z_i - z_j) \\ &< 1 + (z_i - z_j) = z_i + (1 - z_j),\end{aligned}$$

so the total absolute error decreases strictly.

We repeat this exchange whenever such a pair $i \notin S$, $j \in S$ with $z_i > z_j$ exists. Since there are finitely many subsets of cardinality I, the procedure terminates. At termination, every selected fractional part is at least as large as every unselected fractional part. Hence the final selected set consists of I indices having the largest fractional parts. The resulting vector $\widehat{m}$ is therefore an ORIC solution. By transitivity of weak submajorization,

$$e(\widehat{m}) \preceq_{\mathrm{w}} e(m).$$

It remains only to observe that all ORIC solutions have the same multiset of absolute errors. Indeed, let $t = z_{(I)}$ be the cutoff fractional part, and define

$$H = \{i : z_i > t\}, \qquad T = \{i : z_i = t\}, \qquad L = \{i : z_i < t\}.$$

Every ORIC solution rounds every index in H upward, every index in L downward, and exactly $I - |H|$ indices in T upward. The errors contributed by the cutoff tie set are therefore always $I - |H|$ copies of $1 - t$ and $|T| - (I - |H|)$ copies of t. Thus every ORIC solution has the same decreasingly ordered error vector. In particular,

$$e_{(r)}(m^*) = e_{(r)}(\widehat{m}), \qquad r = 1, \ldots, N.$$

It follows that $e(m^*) \preceq_{\mathrm{w}} e(m)$. Finally, if m is not an ORIC solution, then at least one exchange in the steps above is required. Every such exchange strictly decreases the total absolute error. Hence

$$\sum_{i=1}^{N} |x_i - m_i^*| < \sum_{i=1}^{N} |x_i - m_i|.$$

This completes the proof. □

The weak-majorization theorem yields a collection of simultaneous optimality results.

Corollary 3.10 (Universal optimality)**.** *Under the assumptions of Theorem 3.9, let m^* be any ORIC solution. Then the following assertions hold.*

1. *For every $k = 1, \ldots, N$, m^* minimizes the sum of the k largest absolute rounding errors:*

$$\sum_{r=1}^{k} e_{(r)}(m^*) = \min_{m \in \mathcal{F}} \sum_{r=1}^{k} e_{(r)}(m).$$

2. *For every convex nondecreasing function $\psi : [0,\infty) \longrightarrow \mathbb{R}$, m^* minimizes the separable loss $\sum_{i=1}^N \psi\big(|x_i - m_i|\big)$:*

$$\forall m \in \mathcal{F}, \qquad \sum_{i=1}^N \psi\big(|x_i - m_i^*|\big) \le \sum_{i=1}^N \psi\big(|x_i - m_i|\big).$$

3. *m^* minimizes every symmetric, convex, coordinatewise nondecreasing function of the absolute-error vector.*

4. *In particular, m^* simultaneously minimizes*

$$\|x - m\|_q, \qquad 1 \le q \le \infty,$$

and every symmetric gauge norm of the error vector.

Proof. Part 1 is exactly the defining inequality of weak submajorization. We give a direct proof of Part 2. For $v \in \mathbb{R}_+^N$ and $s \ge 0$, define

$$H_s(v) = \sum_{i=1}^N (v_i - s)_+, \qquad (r)_+ = \max\{r, 0\}.$$

If the components of v are arranged in decreasing order, then

$$H_s(v) = \max_{0 \le k \le N} \{T_k(v) - ks\}.$$

Indeed, the maximum is attained when k is the number of components of v that exceed s. Since

$$e(m^*) \preceq_{\mathrm{w}} e(m),$$

we have

$$T_k(e(m^*)) \le T_k(e(m))$$

for every k. Therefore

$$H_s(e(m^*)) = \max_{0 \le k \le N} \{T_k(e(m^*)) - ks\} \le \max_{0 \le k \le N} \{T_k(e(m)) - ks\} = H_s(e(m)).$$

Now suppose first that ψ is a convex, nondecreasing, piecewise-linear function. It can be written in the form

$$\psi(t) = c + \alpha t + \sum_{\ell=1}^J \beta_\ell (t - s_\ell)_+,$$

where $\alpha \ge 0, \beta_\ell \ge 0$. Consequently,

$$\begin{aligned}\sum_{i=1}^N \psi(e_i(m^*)) &= Nc + \alpha T_N(e(m^*)) + \sum_{\ell=1}^J \beta_\ell H_{s_\ell}(e(m^*)) \\ &\le Nc + \alpha T_N(e(m)) + \sum_{\ell=1}^J \beta_\ell H_{s_\ell}(e(m)) = \sum_{i=1}^N \psi(e_i(m)).\end{aligned}$$

For a general convex nondecreasing function ψ, choose an interval $[0, R]$ containing all components of $e(m^*)$ and $e(m)$. Convex nondecreasing piecewise-linear functions obtained by interpolation on increasingly fine partitions converge uniformly to ψ on $[0, R]$. Applying the preceding inequality to these approximations and passing to the limit proves Part 2.

Part 3: majorization property. If $u \preceq_{\mathrm{w}} v$, then $\Phi(u) \leq \Phi(v)$ for every symmetric, convex, coordinatewise nondecreasing function Φ. Applying this result with

$$u = e(m^*), \qquad v = e(m)$$

gives the assertion. For $1 \leq q < \infty$, take $\psi(t) = t^q$ in Part 2. Since the function $r \mapsto r^{1/q}$ is increasing, this gives

$$\|x - m^*\|_q \leq \|x - m\|_q.$$

For $q = \infty$,

$$\|x - m^*\|_\infty = e_{(1)}(m^*) \leq e_{(1)}(m) = \|x - m\|_\infty$$

by Part 1 with $k = 1$. Finally, the monotonicity of symmetric gauge norms under weak submajorization is the Ky Fan dominance principle [18], which gives the last assertion. □

Remark 3.11. Theorem 3.9 strengthens simultaneous L^q-optimality by showing that ORIC minimizes every Ky Fan sum of the absolute-error vector and, more generally, every symmetric convex coordinatewise nondecreasing loss. For example, it shows that ORIC minimizes the largest individual error, the sum of the two largest errors, and, more generally, the sum of the k largest errors for every k, all with the *same* integer rounding.

4 Algorithms for optimal rounding under integer constraints

Proposition 3.1 reduces the complexity of Problem 1 by confining the search to the finite set

$$\prod_{i=1}^{N} \{\lfloor x_i \rfloor, \lceil x_i \rceil\}$$

which can be reparameterized as $\{0, 1\}^N$ i.e. a componentwise rounding problem. But the size of this set grows exponentially with N, so optimization through exhaustive search is not a feasible option, even if N is only moderately large.

4.1 Implementation via sorting

We first exhibit a sorting-based algorithm which exploits the structure of the problem, in particular the sum constraint, to compute an optimal solution in $O(N \log N)$ time. Given the structure of the objective function in Problem 1, the idea is to optimize term by term, controlling for the constraint at each step.

We start by rounding all components downwards and compute the constraint shortfall $I = \sum_{i=1}^{N}(x_i - m_i) \leq N$. If $I = 0$ this means x is already integer valued so one does not need to proceed further. If $I \geq 1$ we need to round upwards exactly I components to meet the constraint. To choose these components optimally, we

1. sort the indices according to decreasing values of fractional part $x_i - \lfloor x_i \rfloor = x_i - m_i$.

2. For a block of equal fractional parts t, order the indices by decreasing integer part when $t < 1/2$, by increasing integer part when $t > 1/2$, and arbitrarily when $t = 1/2$.

3. In the last step, we proceed to round upwards the first I components sorted in this order.

Steps 1) and 2) may be done using a QuickSort algorithm [11].

This yields an algorithm for solving the optimal rounding problem:

Optimal rounding under integer constraints (ORIC**):**

1: Set $\forall i = 1..N, m_i = \lfloor x_i \rfloor$.
2: Compute constraint shortfall $I(x) = \sum_{i=1}^{N}(x_i - m_i) \geq 0$.
3: If $I = 0$ then **end**.
4: Sort indices in decreasing order of fractional part $(x_i - m_i)$:

$$(x_1 - m_1) \geq ... \geq (x_N - m_N) \geq 0.$$

5: **Tie break**: Sort each block of indices with equal fractional parts $0 < t < 1$

- by decreasing integer part if $t < 1/2$,
- by increasing integer part if $t > 1/2$, and
- arbitrarily if $t = 1/2$.

6: **for** $k = 1, \ldots, I$ **do**
7: $\quad m_k = \lceil x_k \rceil$.
8: **end for**
9: **end**

There exists a symmetric version of the algorithm where one initializes with $m_i = \lceil x_i \rceil$ and then needs to round downwards $K - I$ of the components where

$$K = |\{i = 1..N, z_i > 0\}|$$

It is readily verified that the two methods yield the same solutions.

A similar algorithm, based on the sorting of fractional components, but without the sorting step (5), has been used for a long time in the context of proportional seat allocation for representative assemblies with party list voting systems [1], where it is known as the Hare-Niemeyer or 'largest remainder

method'. The focus of the Hare-Niemeyer and related methods is to achieve a 'fair' allocation rather than to minimize a given objective function, so the notion of optimality considered here may or may not be relevant for such applications and more complex considerations apply; we refer to Balinski & Young [1] for a detailed discussion of seat allocation methods in elections. Nevertheless, the largest remainder method is a special case of the method considered here, when all fractional parts are distinct and the analysis of optimality bears many similarities [19].

To quantify the complexity of the algorithm, we first note that

$$I = \sum_{i=1}^{N} (x_i - m_i) \leq N,$$

thus the sorting steps 4 and 5 may be implemented using randomized QuickSort [11] in expected $O(N \log N)$ time [11, 7]. Alternatively, a worst-case $O(N \log N)$ comparison sort gives a deterministic $O(N \log N)$ implementation.

Once the sorting has been done, the rounding of the sorted sequence (Steps 6-8) requires $I \leq N - 1$ operations. So, overall, the complexity is dominated by the sorting step:

Proposition 4.1 (ORIC algorithm: implementation by sorting)**.**
The sorting-based ORIC algorithm returns an ORIC solution and therefore solves Problem 1 for every $q \geq 1$. It has $O(N \log N)$ time complexity and uses $O(N)$ additional space.

Proof. By Proposition 3.2, rounding upward the I coordinates with the largest fractional parts produces an ORIC solution and therefore solves Problem 1 for every $q \geq 1$. Computing a_i, z_i, and I requires $O(N)$ operations. Sorting the N indices by decreasing z_i requires $O(N \log N)$ time, and rounding upward the first I coordinates requires at most N additional operations. Hence the overall running time is $O(N \log N)$, with $O(N)$ additional space.

□

4.2 Deterministic implementation with linear complexity

The implementation described above sorts all N fractional parts. Full sorting is not necessary: it is sufficient to identify the fractional part at the rounding cutoff and partition the coordinates relative to this cutoff. We now use this observation to further reduce the complexity from $O(N \log N)$ to $O(N)$.

We use the following standard selection operation. Given n real numbers and an integer $r \in \{1, \ldots, n\}$, let

$$\text{Select}_r(v_1, \ldots, v_n)$$

denote the r-th largest value among $v_1, \ldots, v_n$. The deterministic median-of-medians selection algorithm computes this order statistic in $O(n)$ worst-case time; see [3].

For each $i = 1, \ldots, N$, write $a_i = \lfloor x_i \rfloor, \qquad z_i = x_i - a_i$, and recall that

$$I = M - \sum_{i=1}^{N} a_i = \sum_{i=1}^{N} z_i \in \mathbb{N}.$$

If $I \geq 1$, define the cutoff fractional part by

$$t = \mathrm{Select}_I(z_1, \ldots, z_N) = z_{(I)}.$$

where $z_{(1)} \geq \cdots \geq z_{(N)}$ are the fractional parts in decreasing order. Partition the indices into

$$H = \{i : z_i > t\}, \qquad T = \{i : z_i = t\}, \qquad L = \{i : z_i < t\}.$$

Every index in H must be rounded upward and every index in L must be rounded downward. The number of indices that must be selected from the cutoff tie set T is $k = I - |H|$. By the definition of $t = z_{(I)}$, $1 \leq k \leq |T|$.

This leads to the following deterministic linear-time implementation:

Deterministic linear-time optimal rounding under integer constraints (Linear-ORIC):

1. Set $m_i = \lfloor x_i \rfloor, \qquad z_i = x_i - m_i, \qquad i = 1, \ldots, N.$
2. Compute the constraint shortfall $I = \sum_{i=1}^{N} z_i$.
3. If $I = 0$, terminate and return $m = x$.
4. Compute the cutoff fractional part $t = \mathrm{Select}_I(z_1, \ldots, z_N)$.
5. Form the index sets
$$H = \{i : z_i > t\}, \qquad T = \{i : z_i = t\}.$$
6. For every $i \in H$, set $m_i = \lceil x_i \rceil$.
7. Compute $k = I - |H|$.
8. Select a subset $C \subseteq T$ with $|C| = k$ as follows:
   - if $t < 1/2$, let C consist of the k indices in T having the largest integer parts a_i;
   - if $t > 1/2$, let C consist of the k indices in T having the smallest integer parts a_i;
   - if $t = 1/2$, choose any k indices from T.

   Ties in integer parts at the secondary cutoff may be resolved arbitrarily.
9. For every $i \in C$, set $m_i = \lceil x_i \rceil$. Return m.

The secondary selection in Step 8 does not require sorting the tie set. For example, when $t < 1/2$, one may determine the k-th largest value among

$$\{a_i : i \in T\}$$

using the same deterministic linear-time selection algorithm, include all indices whose integer parts are strictly larger than this value, and then include sufficiently many indices at equality to obtain exactly k indices. The case $t > 1/2$ is treated analogously using the k-th smallest integer part.

Proposition 4.2 (Deterministic linear-time complexity)**.** *The Linear-ORIC algorithm returns a solution to Problem 1 for every $q \geq 1$. With the secondary selection rule in Step 8, it also returns a solution to Problem 4. The algorithm has deterministic worst-case time complexity $O(N)$ and uses $O(N)$ additional space.*

Proof. If $I = 0$, then

$$\sum_{i=1}^{N} z_i = 0.$$

Since $z_i \geq 0$, this implies $z_i = 0$ for every i, so $x \in \mathbb{N}^N$. Step 3 returns $m = x$, which is the unique solution of Problem 1 and hence also the unique solution of Problem 4. We may therefore assume below that $I \geq 1$. Let $t = z_{(I)}$. Every index $i \in H$ satisfies $z_i > t$, and hence its fractional part is strictly larger than the I-th largest fractional part. Consequently, every such index belongs to the set of the I largest fractional parts and must be rounded upward in every solution characterized by Proposition 3.2.

Similarly, if $i \in L$, then $z_i < t$, so i cannot belong to a set of I indices having the largest fractional parts. Such a coordinate must therefore be rounded downward. The remaining freedom concerns only the cutoff tie set

$$T = \{i : z_i = t\}.$$

Since $|H| < I \leq |H| + |T|$, exactly $k = I - |H|$ indices from T must be rounded upward. Hence the vector returned by the algorithm rounds upward exactly I coordinates having the I largest fractional parts. Proposition 3.2 therefore implies that the returned vector solves Problem 1 for every $q \geq 1$.

If the relative-error tie-breaking rule in Step 8 is used, then among the indices in T the algorithm selects the k largest values of x_i when $t < 1/2$, the k smallest values of x_i when $t > 1/2$, and an arbitrary set when $t = 1/2$. Since all indices in T have the same fractional part, ordering by x_i is equivalent to ordering by $a_i = \lfloor x_i \rfloor$. Corollary 3.4 therefore implies that the returned vector also solves Problem 4.

It remains to establish the complexity bound. Computing a_i, z_i, and I requires $O(N)$ operations. The cutoff $t = z_{(I)}$ can be found in deterministic $O(N)$ time by a linear-time selection algorithm. Forming H, T, and L requires one additional pass through the coordinates and therefore $O(N)$ time.

If a secondary tie-break is required, the relevant order statistic inside T can be found in $O(|T|)$ time, and the selected subset can be constructed in a further $O(|T|)$ pass. Since $|T| \leq N$, this step also requires $O(N)$ time. All remaining assignments require at most N operations. Thus the total deterministic worst-case running time is $O(N)$. Storing the fractional parts and the index sets requires $O(N)$ additional space. □

Remark 4.3. The complexity statements are made in the real-RAM comparison model, in which arithmetic operations and exact comparisons of the input coordinates have unit cost. For floating-point data, equality at the cutoff should be handled using the precision conventions appropriate to the application.

We note further that Linear-ORIC is optimal for large problems:

Corollary 4.4 (Asymptotic optimality)**.** *In the standard RAM model, the worst-case running time $\Theta(N)$ of Linear-ORIC is asymptotically optimal.*

Proof. Returning the full output vector $m = (m_1, \ldots, m_N)$ requires writing N coordinates and therefore takes $\Omega(N)$ time. Proposition 4.2 gives a matching $O(N)$ upper bound. Hence the worst-case running time is $\Theta(N)$. □

Remark 4.5. *If only a solution to Problem 1 is required, the secondary selection in Step 8 is unnecessary: any k indices from the cutoff tie set T may be rounded upward. The relative-error ordering is required only when one also wishes to solve Problem 4.*

5 Comparison with other rounding methods

The sorting-based algorithm described in Section 4.1 has $O(N \log N)$ complexity. The deterministic selection-based implementation of Section 4.2 avoids full sorting and has $O(N)$ worst-case complexity. The identification of the I largest fractional parts is essential and any attempt to bypass it with simpler rounding methods, whether deterministic [26] or randomized [24], may fail to yield the optimal solution. Sorting is one convenient way to perform this step. Using a worst-case $O(N \log N)$ comparison sort yields a deterministic $O(N \log N)$ implementation. Standard randomized QuickSort yields expected $O(N \log N)$ running time.

The asymptotic bias of rounding methods has been studied by Diaconis & Freedman [5] and more recently by Janson [14]. The argument of vanishing asymptotic bias is often used to argue that these methods yield 'unbiased' solutions. We will argue here that, unlike what is suggested by the asymptotic properties, the finite sample solution is in fact systematically biased, in a way that has significant implications for the applications considered in Section 2.4.

5.1 Fractional rounding

A rounding algorithm, often used by default, is *fractional rounding*, which rounds to $\lfloor x \rfloor$ (resp. $\lceil x \rceil$) if $x_i - \lfloor x_i \rfloor \leq \theta$ (resp. $x_i - \lfloor x_i \rfloor > \theta$) where $0 \leq \theta < 1$. $\theta = 1/2$ corresponds to mid-point rounding. Define

$$U_\theta(x) = \{i : z_i > \theta\}, \qquad z_i = x_i - \lfloor x_i \rfloor.$$

Threshold rounding is feasible precisely when $|U_\theta(x)| = I$. If $z_{(I)} > z_{(I+1)}$, any threshold satisfying

$$z_{(I+1)} \leq \theta < z_{(I)}$$

produces the ORIC solution. If $z_{(I)} = z_{(I+1)}$, no pure threshold can in general select exactly I coordinates; an additional tie-breaking rule is required.

So, a fractional rounding rule may fail to yield the optimal solution to Problem 1 for some $x \in \mathbb{R}_+^N$. Consequently, no fixed pure-threshold rounding rule is feasible and optimal for every admissible input x.

5.2 Independent randomized rounding

We now compare the algorithm described above with randomized rounding [24], in which each x_i is rounded up with probability $p_i = x_i - \lfloor x_i \rfloor$. Randomized rounding thus yields a solution given by

$$R_i = \lfloor x_i \rfloor + U_i, \qquad U_i \sim \text{Bernoulli}(p_i), \qquad p_i = x_i - \lfloor x_i \rfloor. \tag{11}$$

$$\mathbb{P}(U = u) = \prod_{i=1}^{N} p_i^{u_i}(1-p_i)^{1-u_i}, \qquad u \in \{0,1\}^N.$$

Assume, without loss of generality, that $1 > p_1 \geq ... \geq p_N \geq 0$. Let $I = \sum_{i=1}^{N}(x_i - \lfloor x_i \rfloor)$. Then the optimal solution of Problem 1 corresponds to

$$m^* = (\underbrace{1, 1, ..., 1}_{I}, 0, ...0) + \lfloor x \rfloor \tag{12}$$

The probability of obtaining this optimum by randomized rounding is

$$\mathbb{P}(R = m^*) = \prod_{i=1}^{I} p_i \prod_{i=I+1}^{N} (1-p_i)$$

If the solution is unique, the probability that randomized rounding gives an incorrect solution is

$$\mathbb{P}(R \neq m^*) = 1 - \prod_{i=1}^{I} p_i \prod_{i=I+1}^{N} (1-p_i)$$

This probability can be higher than 50%, as the following example shows.

Example 5.1. *Let $N = 3$ and $x - \lfloor x \rfloor = (0.4, 0.35, 0.25)$. The optimal solution is to round up 0.4 and round down the other components to zero. Using the formula above the probability of finding the optimum is*

$$0.4 \times 0.65 \times 0.75 = 0.195$$

so the probability of an incorrect result is 0.805.

Since $\sum p_i = I$, when N is large the normalized discrepancy goes to zero in probability by the law of large numbers. Indeed,

$$\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}(U_i - p_i)\right)^2\right] = \frac{1}{N^2}\sum_{i=1}^{N} p_i(1-p_i) \leq \frac{1}{4N}.$$

Hence

$$\frac{1}{N}\left(\sum_{i=1}^{N} R_i - M\right) = \frac{1}{N}\sum_{i=1}^{N}(U_i - p_i) \longrightarrow 0$$

in L^2, and therefore in probability. However, the probability of the constraint being satisfied may not go to one. For example if $p_i = 1/2$ and N is even then

$$\mathbb{P}(\sum_{i=1}^{N} U_i = \frac{N}{2}) = \binom{N}{\frac{N}{2}} 2^{-N} \sim \sqrt{\frac{2}{\pi N}} \overset{N\to\infty}{\to} 0.$$

Independent randomized rounding (11) generally yields a *biased* estimator of the optimal integer vector m^*. Under a strict cutoff, the sign of $\mathbb{E}[R_i] - m_i^*$ depends on whether the coordinate is rounded upward or downward by ORIC:

Proposition 5.1 (Independent randomized rounding)**.** *Let $x \notin \mathbb{Z}^N$ with $\sum_i x_i = M \in \mathbb{Z}$ and $p_i = x_i - \lfloor x_i \rfloor$. Assume without loss of generality*

$$1 > p_1 = x_1 - \lfloor x_1 \rfloor \geq ... \geq p_N = x_N - \lfloor x_N \rfloor \geq 0.$$

Define

$$I = \sum_{i=1}^{N} p_i = M - \sum_{i=1}^{N} \lfloor x_i \rfloor \in \mathbb{N}$$

Then $1 \leq I \leq N-1$. Let $U_i \sim$Bernoulli(p_i) be independent variables with $\mathbb{P}(U_i = 1) = p_i$ and define the randomized rounding scheme $R_i = \lfloor x_i \rfloor + U_i$. Then

1. *Randomized rounding satisfies the integer constraint in expectation*

$$E\left(\sum_{i=1}^{N} R_i\right) = M$$

2. *Assume $p_I > p_{I+1}$ so that the optimal rounding is unique. Then $m_i^* = \lfloor x_i \rfloor + 1_{i \leq I}$ and*

$$E(R_i) - m_i^* = -(1 - p_i)1_{i \leq I} + p_i \; 1_{i > I}$$

3. *The probability that randomized rounding produces the optimal vector is*

$$\mathbb{P}(R = m^*) = \prod_{i=1}^{I} p_i \prod_{i=I+1}^{N} (1 - p_i) \tag{13}$$

The main justification often advanced for randomized rounding is that 'it yields an unbiased estimator'. However R_i is an unbiased estimator of x_i, not the desired integer solution m_i^*.

As Proposition 5.1 shows, randomized rounding preserves integer constraints in expectation only and, in general, neither satisfies the sum constraint almost surely nor returns the optimal integer rounding. It also leads to a systematic upward bias in components whose fractional part is small.

5.3 Dependent randomized rounding

Dependent randomized rounding addresses an important limitation of independent randomized rounding by introducing dependence between the coordinatewise rounding decisions. cardinality-preserving dependent-rounding scheme of Gandhi et al. [8] is particularly appealing as it enables to preserve the sum constraint. Applied to the fractional parts

$$z_i = x_i - \lfloor x_i \rfloor, \qquad \sum_{i=1}^{N} z_i = I \in \mathbb{N},$$

the cardinality-preserving dependent-rounding scheme of Gandhi et al. [8] generates binary random variables $U_1, \ldots, U_N$ satisfying

$$\mathbb{P}(U_i = 1) = z_i, \quad i = 1, \ldots, N, \quad \text{and} \qquad \sum_{i=1}^{N} U_i = I \qquad \text{with probability one.}$$

The rounded vector $R_i = \lfloor x_i \rfloor + U_i$ therefore satisfies

$$\sum_{i=1}^{N} R_i = M \qquad \text{with probability one,}$$

while preserving the coordinatewise marginals. The resulting binary variables also satisfy negative-correlation properties which are useful for obtaining concentration bounds for other linear combinations of the rounded coordinates.

Dependent rounding therefore improves substantially on independent randomized rounding with respect to feasibility: the sum constraint is satisfied exactly, rather than only in expectation. It nevertheless addresses a different criterion from Problem 1.

Independent and dependent randomized rounding have the same expected separable L^q loss when they use the same coordinatewise marginals, although dependent rounding satisfies the sum constraint almost surely. But preservation of the marginals does not imply that each realized integer vector minimizes the rounding error. Indeed, for every $q \geq 1$,

$$\mathbb{E}\left[\sum_{i=1}^{N} |x_i - R_i|^q\right] = \sum_{i=1}^{N} \left[(1 - z_i) z_i^q + z_i (1 - z_i)^q\right].$$

This expression depends only on the marginal probabilities z_i, and not on the dependence structure among the variables U_i. Since every realization of dependent rounding is a feasible floor-or-ceiling rounding, Proposition 3.2 gives

$$\sum_{i=1}^{N} |x_i - m_i^*|^q \leq \sum_{i=1}^{N} |x_i - R_i|^q \qquad \text{with probability one,}$$

where m^* is any ORIC solution. Consequently,

$$\sum_{i=1}^{N} |x_i - m_i^*|^q \leq \mathbb{E}\left[\sum_{i=1}^{N} |x_i - R_i|^q\right].$$

When the ORIC solution is unique and $I \geq 1$, this inequality is strict: marginal preservation requires each selected coordinate (i) to be rounded downward with probability $1 - z_i > 0$, so dependent rounding returns a nonoptimal feasible vector with positive probability.

For example, consider the fractional parts

$$(z_1, z_2, z_3) = (0.4, 0.35, 0.25), \qquad I = 1.$$

Any cardinality-preserving dependent rounding with these marginals rounds upward exactly one coordinate, choosing coordinates $(1, 2, 3)$ with probabilities $(0.4, 0.35, 0.25)$, respectively. The ORIC solution rounds the first coordinate upward and is therefore obtained with probability 0.4. The three possible total absolute errors are, respectively,

$$1.2, \qquad 1.3, \qquad 1.5,$$

so dependent rounding has expected total absolute error

$$0.4 \times 1.2 + 0.35 \times 1.3 + 0.25 \times 1.5 = 1.31,$$

whereas ORIC attains the minimum value (1.2) deterministically.

Thus, dependent rounding and ORIC provide complementary guarantees. Dependent rounding is appropriate when exact feasibility must be combined with prescribed marginal probabilities, repeated-allocation fairness, or probabilistic concentration properties. ORIC is appropriate when the objective is to compute, in a single realization, an exactly feasible integer vector having minimum rounding error.

6 Matroid-constrained optimal rounding

The sum constraint in 1 is equivalent to choosing I elements from N, meaning that the upward-rounded sets are the bases of the uniform matroid $U_{I,N}$. This suggests an extension of ORIC to matroid-constrained optimal rounding.

We first recall some definitions [15]. A *matroid* is a pair $\mathcal{M} = (E, \mathcal{I})$, where E is a finite ground set and $\mathcal{I} \subseteq 2^E$ is a nonempty family of subsets, called the independent sets, such that:

1. if $A \in \mathcal{I}$ and $B \subseteq A$, then $B \in \mathcal{I}$;
2. if $A, B \in \mathcal{I}$ and $|A| < |B|$, then there exists $j \in B \setminus A$ such that $A \cup \{j\} \in \mathcal{I}$.

A maximal independent set is called a *basis*. The set of bases of $\mathcal{M}$ is denoted by $\mathcal{B}(\mathcal{M})$. All bases have the same cardinality, called the rank of $\mathcal{M}$ and denoted by $r(\mathcal{M})$.

The uniform matroid $U_{r,n}$ on a ground set of cardinality n is the matroid whose independent sets are the subsets of cardinality at most r. Its bases are therefore the subsets of cardinality exactly r.

Let $x = (x_1, \dots, x_N) \in \mathbb{R}_+^N$ satisfy $\sum_{i=1}^N x_i = M \in \mathbb{N}$. For each i, write $a_i = \lfloor x_i \rfloor$ and $z_i = x_i - a_i \in [0, 1)$. As before,

$$I = M - \sum_{i=1}^N a_i = \sum_{i=1}^N z_i \in \mathbb{N}. \tag{14}$$

Coordinates for which $z_i = 0$ are already integer-valued and are fixed in every componentwise rounding. Define

$$E_x = \{i \in \{1, \dots, N\} : z_i > 0\}. \tag{15}$$

Let $\mathcal{M} = (E_x, \mathcal{I})$ be a matroid of rank I. For each basis $B \in \mathcal{B}(\mathcal{M})$, define $m^B \in \mathbb{N}^N$ by

$$m_i^B = a_i + \mathbf{1}_{\{i \in B\}}, \qquad i = 1, \dots, N. \tag{16}$$

Then $m_i^B \in \{\lfloor x_i \rfloor, \lceil x_i \rceil\}$ for every i, and

$$\sum_{i=1}^N m_i^B = \sum_{i=1}^N a_i + |B| = \sum_{i=1}^N a_i + I = M. \tag{17}$$

Thus every basis gives a feasible sum-preserving rounding.

For $B \in \mathcal{B}(\mathcal{M})$, define the absolute-error vector

$$e(B) = \big(|x_1 - m_1^B|, \dots, |x_N - m_N^B|\big). \tag{18}$$

Its components satisfy

$$e_i(B) = \begin{cases} 1 - z_i, & i \in B, \\ z_i, & i \notin B. \end{cases} \tag{19}$$

For $B \subseteq E_x$, define its fractional-part weight by

$$w(B) = \sum_{i \in B} z_i. \tag{20}$$

Problem 5 (Matroid-constrained optimal rounding)**.** *Let $q \geq 1$. Given x and a rank-I matroid $\mathcal{M}$ on E_x, find a basis $B \in \mathcal{B}(\mathcal{M})$ minimizing*

$$\|x - m^B\|_q = \left(\sum_{i=1}^N e_i(B)^q \right)^{1/q}. \tag{21}$$

For $q = \infty$, the objective is $\|x - m^B\|_\infty = \max_i e_i(B)$.

When $\mathcal{M} = U_{I,|E_x|}$, every subset of E_x of cardinality I is a basis, so Problem 5 reduces to Problem 1.

Recall that for $u, v \in \mathbb{R}_+^N$, the relation $u \preceq_w v$ means that

$$\sum_{r=1}^{k} u_{(r)} \leq \sum_{r=1}^{k} v_{(r)}, \qquad k = 1, \ldots, N, \tag{22}$$

where $u_{(1)} \geq \cdots \geq u_{(N)}$ and $v_{(1)} \geq \cdots \geq v_{(N)}$ are the decreasing rearrangements of u and v.

Theorem 6.1 (Matroid-constrained universal rounding)**.** *Let $\mathcal{M}$ be a rank-I matroid on E_x, and let*

$$B^\star \in \operatorname*{arg\,max}_{B \in \mathcal{B}(\mathcal{M})} w(B) \tag{23}$$

be any maximum-weight basis for the weights z_i. Then, for every $B \in \mathcal{B}(\mathcal{M})$,

$$e(B^\star) \preceq_w e(B). \tag{24}$$

Equivalently,

$$\sum_{r=1}^{k} e_{(r)}(B^\star) \leq \sum_{r=1}^{k} e_{(r)}(B), \qquad k = 1, \ldots, N. \tag{25}$$

All maximum-weight bases have the same decreasingly ordered absolute-error vector. Moreover, for every basis B,

$$\sum_{i=1}^{N} e_i(B) = 2I - 2w(B). \tag{26}$$

Consequently, a basis minimizes the total absolute error if and only if it is a maximum-weight basis.

Every maximum-weight basis simultaneously minimizes $\|x - m^B\|_q$ for all $1 \leq q \leq \infty$. More generally, it minimizes every Ky Fan sum of the absolute-error vector, every separable loss $\sum_{i=1}^{N} \psi(e_i(B))$ with ψ convex and nondecreasing, every symmetric convex coordinatewise nondecreasing function of $e(B)$, and every symmetric gauge norm of $e(B)$.

Proof. If $I = 0$, then $z_i = 0$ for all i, the unique basis is the empty set, and the result is immediate. Assume henceforth that $I \geq 1$.

Order the elements of E_x as $i_1, \ldots, i_n$, where $n = |E_x|$, so that

$$z_{i_1} \geq z_{i_2} \geq \cdots \geq z_{i_n}. \tag{27}$$

Resolve ties arbitrarily. Apply the matroid greedy algorithm in this order: begin with the empty set and add i_j whenever doing so preserves independence. Let $G = \{g_1, \ldots, g_I\}$ be the resulting basis, with the elements listed in the order in which they were accepted.

Fix an arbitrary basis B. We construct bases $B_0, \dots, B_I$ such that $B_0 = B$, $B_I = G$, $\{g_1, \dots, g_k\} \subseteq B_k$, and $e(B_k) \preceq_w e(B_{k-1})$ for each k.

Suppose that B_{k-1} has been constructed. If $g_k \in B_{k-1}$, set $B_k = B_{k-1}$. Otherwise, $B_{k-1} \cup \{g_k\}$ contains a unique circuit C_k. Since $\{g_1, \dots, g_k\}$ is independent, there exists

$$f_k \in C_k \setminus \{g_1, \dots, g_k\}. \tag{28}$$

Then

$$B_k = \big(B_{k-1} \setminus \{f_k\}\big) \cup \{g_k\} \tag{29}$$

is a basis containing $g_1, \dots, g_k$.

We claim that $z_{g_k} \geq z_{f_k}$. Indeed, if f_k had been inspected before g_k, then the greedy algorithm would have rejected f_k, because f_k is not one of $g_1, \dots, g_{k-1}$. At that time the accepted set was exactly $\{g_1, \dots, g_{k-1}\}$, so rejection would imply that $\{g_1, \dots, g_{k-1}, f_k\}$ is dependent. This is impossible because that set is contained in the basis B_{k-1}. Hence f_k is inspected no earlier than g_k, and therefore $z_{g_k} \geq z_{f_k}$.

Set $a = z_{g_k}$ and $b = z_{f_k}$, so that $0 < b \leq a < 1$. Before the exchange, the errors in coordinates g_k and f_k are $(a, 1-b)$. After the exchange, they are $(1-a, b)$. Since $b \leq a$ and $1 - a \leq 1 - b$,

$$\max\{1-a, b\} \leq \max\{a, 1-b\}. \tag{30}$$

Also,

$$(1-a) + b = 1 - (a-b) \leq 1 + (a-b) = a + (1-b). \tag{31}$$

Thus

$$(1-a, b) \preceq_w (a, 1-b). \tag{32}$$

After adjoining the unchanged coordinates, the embedding property of weak submajorization gives $e(B_k) \preceq_w e(B_{k-1})$.

Iterating the construction yields

$$e(G) = e(B_I) \preceq_w e(B_{I-1}) \preceq_w \cdots \preceq_w e(B_0) = e(B). \tag{33}$$

Therefore $e(G) \preceq_w e(B)$ for every basis B.

For any basis B, using $\sum_i z_i = I$ and $|B| = I$, we obtain

$$\begin{aligned} \sum_{i=1}^{N} e_i(B) \quad &= \textstyle\sum_{i \notin B} z_i + \sum_{i \in B}(1 - z_i) = \sum_{i=1}^{N} z_i + \sum_{i \in B}(1 - 2z_i) \\ &= I + |B| - 2\textstyle\sum_{i \in B} z_i = 2I - 2w(B). \end{aligned} \tag{34}$$

Taking $k = N$ in $e(G) \preceq_w e(B)$ and using (34) shows that $w(G) \geq w(B)$. Hence G is a maximum-weight basis.

Now let $B^\star$ be any maximum-weight basis. Apply the preceding transformation to $B^\star$. Each exchange replaces an element of weight b by one of weight $a \geq b$. Since $w(B^\star) = w(G)$, the sum of all weight increases is zero. Therefore

every nontrivial exchange satisfies $a = b$. In such an exchange, the affected error pair changes from $(a, 1-a)$ to $(1-a, a)$, so its multiset is unchanged. Consequently, $e(B^\star)$ and $e(G)$ have the same decreasing rearrangement. Since $e(G) \preceq_w e(B)$ for every basis B, we obtain $e(B^\star) \preceq_w e(B)$ for every basis B.

Identity (26) shows that minimizing total absolute error is equivalent to maximizing $w(B)$. Finally, all the universal optimality statements follow from standard properties of weak submajorization. In particular, taking $\psi(t) = t^q$ gives simultaneous L^q-optimality for $1 \le q < \infty$, while the case $q = \infty$ follows from the inequality for the largest component. The remaining assertions follow from the monotonicity of symmetric convex coordinatewise nondecreasing functions and of symmetric gauge norms under weak submajorization. □

Corollary 6.2 (Greedy implementation)**.** *Process the indices in nonincreasing order of z_i and add an index whenever doing so preserves matroid independence. The selected set is a maximum-weight basis and gives a universally optimal matroid-constrained rounding.*

Remark 6.3 (Recovery of ORIC)**.** *If $\mathcal{M} = U_{I,|E_x|}$, the greedy algorithm selects the I largest fractional parts. Theorem 6.1 therefore reduces to the weak-majorization optimality theorem for ORIC.*

Remark 6.4 (Computational complexity)**.** *Suppose an independence test for $\mathcal{M}$ takes time T_{ind}. Sorting the fractional parts takes $O(N \log N)$ time, and the greedy algorithm performs at most N independence tests. The total running time is therefore $O(N \log N + N T_{\mathrm{ind}})$.*

References

[1] M. Balinski and H. P. Young, *Fair representation: Meeting the ideal of One man, One Vote*, Yale University Press, 1982.

[2] T. Berthold, *RENS: The optimal rounding*, *Mathematical Programming Computation*, vol. 6, pp. 33–54, 2014.

[3] M. Blum, R. W. Floyd, V. Pratt, R. L. Rivest, and R. E. Tarjan, *Time bounds for selection*, *Journal of Computer and System Sciences*, vol. 7, no. 4, 448–461, 1973.

[4] L. H. Cox and L. R. Ernst, *Controlled rounding*, INFOR, 20 (1982), pp. 423–432.

[5] P. Diaconis and D. Freedman, *On rounding percentages*, Journal of the American Statistical Association, 74 (1979), pp. 359–364.

[6] A. Federgruen and H. Groenevelt, *The greedy procedure for resource allocation problems: Necessary and sufficient conditions for optimality*, Operations Research, 34 (1986), pp. 909–918.

[7] M. L. Fredman, *An intuitive and simple bounding argument for Quicksort*, Information Processing Letters, 114 (2014), pp. 137 – 139.

[8] R. Gandhi, S. Khuller, S. Parthasarathy, and A. Srinivasan, *Dependent rounding and its applications to approximation algorithms*, Journal of the ACM, 53 (2006), pp. 324–360.

[9] G. R. Grimmett, *European apportionment via the Cambridge compromise*, Mathematical Social Sciences, 63 (2012), pp. 68 – 73.

[10] M. Grötschel, L. Lovasz, and A. Schrijver, *Geometric Algorithms and Combinatorial Optimization*, Springer, 1993.

[11] C. A. R. Hoare, *Quicksort*, The Computer Journal, 5 (1962), pp. 10–16.

[12] D. S. Hochbaum and J. G. Shanthikumar, *Convex separable optimization is not much harder than linear optimization*, Journal of the ACM, 37 (1990), pp. 843–862.

[13] T. Ibaraki and N. Katoh, *Resource allocation problems: Algorithmic approaches*, MIT Press, Cambridge, MA, 1988.

[14] S. Janson, *Asymptotic bias of some election methods*, Annals of Operations Research, 215 (2014), pp. 89–136.

[15] B. Korte and J. Vygen, *Combinatorial optimization*, Springer, 2002.

[16] D. E. Knuth, *Two-way rounding*, SIAM Journal on Discrete Mathematics, 8 (1995), pp. 281–290.

[17] J. K. Lenstra, D. B. Shmoys, and E. Tardos, *Approximation algorithms for scheduling unrelated parallel machines*, Mathematical Programming, 46 (1990), pp. 259–271.

[18] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and Its Applications*, 2nd ed., Springer, 2011.

[19] H. F. Niemeyer and A. C. Niemeyer, *Apportionment methods*, Math. Social Sci., 56 (2008), pp. 240–253.

[20] G. Pólya, *Proportionalwahl und Wahrscheinlichkeitsrechnung*, Zeitschrift für die gesamte Staatswissenschaft/Journal of Institutional and Theoretical Economics, (1919), pp. 297–322.

[21] G. Pólya, *Sur la représentation proportionnelle en matière électorale*, L'Enseignement Mathématique, 20 (1919), 355–379.

[22] C. Potts, *Analysis of a linear programming heuristic for scheduling unrelated parallel machines*, Discrete Appl. Math., 10 (1985), 155–164.

[23] F. Pukelsheim, *Proportional representation: Apportionment methods and their applications*, Springer, Cham, 2014.

[24] P. Raghavan and C. D. Thompson, *Randomized rounding: a technique for provably good algorithms and algorithmic proofs*, Combinatorica, 7 (1987), 365–374.

[25] A. S. Schulz and M. Skutella, *Scheduling unrelated machines by randomized rounding*, SIAM Journal on Discrete Mathematics, 15 (2002), pp. 450–469.

[26] D. P. Williamson and D. B. Shmoys, *The design of approximation algorithms*, Cambridge University Press, 2011.