Turnpike with Uncertain Measurements: Triangle-Equality Integer Programming with a Deterministic Recovery Guarantee

C. S. Elder
Ray and Stephanie Lane Computational Biology Department,
Carnegie Mellon University, Pittsburgh, PA, USA

Guillaume Marçais
Ray and Stephanie Lane Computational Biology Department,
Carnegie Mellon University, Pittsburgh, PA, USA

Carl Kingsford
Ray and Stephanie Lane Computational Biology Department,
Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

We study the Turnpike problem with uncertain measurements: reconstructing a one-dimensional point set from an unlabeled multiset of pairwise distances under bounded noise and rounding. We give a combinatorial characterization of realizability via a multi-matching that labels interval indices by distinct distance values while satisfying all triangle equalities. This yields an integer linear program (ILP) based on the triangle equality whose constraint structure depends only on the two-partition set $\mathcal{P}_y = \{(r, s, t) : y_r + y_s = y_t,\ (r \neq s \text{ or } \mu_r \geq 2)\}$ and a natural linear-programming (LP) relaxation with $\{0, 1\}$-coefficient constraints. Integral solutions certify realizability and output an explicit assignment matrix, enabling a modular *assignment-first, regression-second* pipeline for downstream coordinate estimation. Under bounded noise followed by rounding, we prove deterministic separation conditions under which distinct rounded values do not collide and $\mathcal{P}_y$ is recovered exactly, so the ILP/LP receives the same combinatorial input as in the noiseless case. Experiments illustrate integrality behavior and degradation outside the provable regime.

2012 ACM Subject Classification Theory of computation → Design and analysis of algorithms; Mathematics of computing → Discrete optimization; Applied computing → Bioinformatics

Keywords and phrases Turnpike problem, partial digest, integer linear programming, LP relaxation, unlabeled distances, partitions, bounded noise

Digital Object Identifier 10.4230/LIPIcs.WABI.2026.20

Related Version *Full Version*: https://arxiv.org/abs/2603.18283

Funding This work was supported in part by the US National Science Foundation [III-2232121] and the US National Institutes of Health [R01HG012470].

1 Introduction

The *Turnpike problem* is a classical inverse problem: reconstruct a one-dimensional point set from an unlabeled multiset of pairwise distances. The origin story imagines a highway: we are given the multiset of all pairwise distances between exits, but the exit mile markers themselves have been lost. Can we recover a configuration of exits that is consistent with the observed distances?

Formally, let $x \in \mathbb{R}^n$ have strictly increasing coordinates. The *difference multiset* associated with x is $\Delta x := \{x_j - x_i : 1 \leq i < j \leq n\}$, which is a multiset containing $m := \binom{n}{2}$ positive reals. An instance of Turnpike is a multiset Y of m positive reals, and the decision problem asks whether there exists an x with $\Delta x = Y$.

26th International Conference on Algorithms for Bioinformatics (WABI 2026).
Editors: Nadia El-Mabrouk and Fabio Vandin; Article No. 20; pp. 20:1–20:20
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

A multiset Y is represented by a pair (y, μ) such that y lists the values contained in Y in strictly decreasing order $y_1 > \cdots > y_{m'} > 0$ where the multiplicity vector $\mu \in \mathbb{N}_{++}^{m'}$ records the number of occurrences, i.e., that a value y_k is present $\mu_k \geq 1$ times. We say Y is *realizable* when $Y = \Delta x$ for some $x \in \mathbb{R}^n$.

The Turnpike problem is equivalent to the Partial Digest problem from restriction site mapping in computational biology [1, 19] and is closely related to distance-geometry reconstruction tasks in crystallography, tandem mass spectrometry, and molecular error-correcting codes [7, 10, 12]. In restriction mapping, for example, a single-enzyme partial digestion produces fragments whose lengths coincide with all pairwise distances between restriction sites; the multiset Y records fragment lengths, and a solution x specifies restriction site locations along the strand (up to translation and reflection). The same combinatorial core occurs in tandem mass spectrometry, where the prefix masses of a peptide's backbone-fragmentation ions form a point set whose unlabeled pairwise differences are observed, and recovering the ordered residues from that multiset is again a Turnpike-type reconstruction [7, 15].

Exact recovery from error-free distances is neither known to be NP-complete nor to admit a polynomial-time algorithm [3]. A backtracking algorithm with worst-case exponential runtime but polynomial expected runtime under absolutely-continuous input models is known [18, 19]. Many additional approaches have been proposed, including polynomial factorization methods for exact instances, pruning strategies, and hardness and counterexample constructions for related variants [2, 3, 5, 8, 9, 13, 14, 16, 17, 20, 21].

Noisy Turnpike and measurement uncertainty

Experimental pipelines that produce Turnpike-like instances (partial digestion, tandem mass spectrometry, polymer codes) inevitably include measurement uncertainty and missing or duplicated observations. Allowing bounded additive or multiplicative errors makes the corresponding *bounded-error Turnpike* (formalized in Section 2.5) and the related circular Beltway problem strongly NP-hard [4, 6, 12], and backtracking approaches routinely explore exponentially many states even on non-adversarial inputs [4, 6, 19].

These noisy variants are strongly NP-hard, so we do not aim for an exact, polynomial-time algorithm. Instead, we focus on formulations whose integral solutions certify realizability and whose fractional solutions still retain structured combinatorial information, complementing the best-fit reconstruction solvers of prior work [6].

1.1 This work: a triangle-based assignment viewpoint

We develop a complementary viewpoint, treating unlabeled distances as a combinatorial *labeling* problem first and a regression problem second. The key object is a *ruler*: an indexed pairwise-difference table in which entry (i, j) stores $x_j - x_i$. Unlike the multiset Δx, a ruler retains interval indices, and the identity $x_k - x_i = (x_j - x_i) + (x_k - x_j)$ becomes a family of linear consistency constraints (the *triangle equalities*). Realizability of Y can therefore be rephrased as the existence of an assignment of distance values to interval indices that satisfies triangle equalities on every triple $i < j < k$.

This lens leads first to a natural mixed-integer linear programming (MILP) formulation for exact Turnpike, and then to a coordinate-free integer linear programming (ILP) formulation using only combinatorial information. A distinctive feature of this formulation is that it outputs an *assignment matrix* $\widehat{P}$ (a multi-matching between intervals and distinct distance values). A coordinate estimate $\widehat{x}$ can then be computed afterward by solving an independent regression problem with any preferred numeric precision. This makes permutation metrics over $\widehat{P}$ the natural choice for primary evaluation, with coordinate-wise metrics treated as secondary.

In summary, we build a triangle-equality ILP for exact Turnpike, study its LP relaxation as a certificate/diagnostic layer, show that the formulation extends to the uncertain setting under deterministic separation conditions, and empirically validate that performance tracks this regime and degrades rapidly outside it.

Main results and algorithmic viewpoint

Our results fall into four parts.

- **Triangle feasibility as realizability.** Realizability of Y is equivalent to the existence of a multi-matching between interval indices and distinct distance values whose induced ruler satisfies all triangle equalities (Lemma 6), yielding a natural MILP (Proposition 8).
- **A coordinate-free triangle ILP/LP parameterized by two-partitions.** Eliminating coordinate variables and encoding triangle consistency through the two-partition set $\mathcal{P}_y = \{(r,s,t) : y_r + y_s = y_t,\ (r \neq s \text{ or } \mu_r \geq 2)\}$ yields the triangle ILP (Proposition 10) and its LP relaxation (System 11); $\mathcal{P}_y$ is built in $O(m'^2)$ time (Algorithm 1). With the basis reduction of Lemma 13, the formulation has $O(n^4)$ assignment and $O(n^6)$ triangle variables. The role is certificate generation, not best-fit regression.
- **Deterministic robustness under bounded noise and rounding.** Under measurement radius ε, rounding radius R, and the separation condition $\mathrm{gap}_\star > 6(\varepsilon + R/2)$ (with $\mathrm{sep}_y > 2(\varepsilon + R/2)$ guarding distinct-value collisions), the rounded data recover the exact two-partition set $\mathcal{P}_y$ (Theorem 15), so the triangle formulation operates on the noiseless combinatorial input.
- **Empirical evidence and assignment-first evaluation.** Experiments on synthetic and partial-digest instances track integrality and recovery primarily through assignment-level metrics (labelings and permutation distances) rather than coordinate regression. Recovery degrades rapidly in the (ε, R) phase diagram (Figure 2) once Theorem 15's separation regime is violated.

Limitations

Our deterministic guarantee assumes bounded noise followed by rounding and does not cover missing or duplicated distances. The separation condition is moreover stated through the noiseless quantities sep_y and $\mathrm{gap}_\star$, which are properties of the unknown ground truth rather than observable inputs; in practice we replace them by the data-driven calibration of Section 3.7. Although the formulations use polynomially many constraints and variables in n and m', the $O(n^6)$ triangle variables make full ILP/LP sweeps impractical beyond moderate sizes (we sweep up to $n = 30$), and we leave large-scale separation/enforcement schemes for future work. We also do not characterize when the LP relaxation is integral: integrality is reported empirically and a supporting theory is left open. Finally, our guarantee addresses additive noise with rounding; multiplicative and signal-dependent error models would require a different separation analysis.

2 Preliminaries

This section fixes notation and introduces the objects used throughout: distance multisets and the index sets they live over (Section 2.1), intervals and rulers (Section 2.2), the multi-matching viewpoint that converts realizability into an assignment-feasibility question

(Section 2.3), the corresponding MILP (Section 2.4), and the (ε, R) noise-and-rounding model used later for deterministic recovery guarantees (Section 2.5). Symbols are introduced as they appear and are collected for reference in the notation table of Appendix C; the two-partition set $\mathcal{P}_y$, the central object of the triangle ILP/LP, is defined in Section 3.1.

2.1 Notation and multisets

We write $\mathbb{R}$, $\mathbb{Z}$, and $\mathbb{N}$ for the reals, integers, and naturals, respectively; subscripts indicate standard restrictions (e.g., $\mathbb{R}_+$ and $\mathbb{R}_{++}$ for nonnegative and positive reals). For an integer $n \geq 1$ we use $[n] = \{1, 2, \ldots, n\}$.

A numeric multiset is represented as an ordered pair $Y = (y, \mu)$ where $y \in \mathbb{R}^{m'}$ lists the distinct values in strictly decreasing order $y_1 > \cdots > y_{m'} > 0$, and $\mu \in \mathbb{N}_{++}^{m'}$ records the multiplicity of each value. The total cardinality (counting multiplicity) is $m = \sum_{r=1}^{m'} \mu_r$.

As in Section 1, for $x \in \mathbb{R}^n$ with strictly increasing entries we write $\Delta x = \{x_j - x_i : 1 \leq i < j \leq n\}$ for the multiset of pairwise distances. When $\Delta x = (y, \mu)$, the total multiplicity satisfies $m = \binom{n}{2}$.

▶ **Example 1.** For $x = (0, 2, 5, 9)$, the distances are $\{2, 3, 4, 5, 7, 9\}$, so $m' = m = 6$ and $\mu = \mathbf{1}$. For $x = (0, 2, 4, 6)$, the distances are $\{2, 2, 2, 4, 4, 6\}$, so $y = (6, 4, 2)$ and $\mu = (1, 2, 3)$.

2.2 Intervals, refinements, and rulers

We index unknown distances by their endpoints. A *monotone interval* is an ordered pair in $\mathcal{I}_n \coloneqq \{(i, j) \in [n] \times [n] : i < j\}$; a *monotone refinement* is an ordered triple in $\mathcal{T}_n \coloneqq \{(i, j, k) \in [n]^3 : i < j < k\}$. When unambiguous, we shorten (i, j) to ij and (i, j, k) to ijk.

▶ **Definition 2.** *A* ruler *on $[n]$ is a matrix $\rho \in \mathbb{R}^{[n]\times[n]}$ for which there exists $x \in \mathbb{R}^n$ such that $\rho_{ij} = x_j - x_i$ for all $i, j \in [n]$. We call x a* realization *of ρ. The ruler is* monotone *if $\rho_{ij} > 0$ for all $ij \in \mathcal{I}_n$.*

Every ruler satisfies skew-symmetry ($\rho_{ij} = -\rho_{ji}$) and the triangle equalities $\rho_{ik} = \rho_{ij} + \rho_{jk}$. The converse also holds.

▶ **Lemma 3** (Triangle equalities characterize rulers)**.** *A matrix $\rho \in \mathbb{R}^{[n]\times[n]}$ is a ruler if and only if it satisfies $\rho_{ik} = \rho_{ij} + \rho_{jk}$ for all $i, j, k \in [n]$.*

Proof. If $\rho_{ij} = x_j - x_i$ for some x, then $\rho_{ik} = (x_k - x_i) = (x_j - x_i) + (x_k - x_j) = \rho_{ij} + \rho_{jk}$.

Conversely, suppose ρ satisfies the triangle equalities for all triples. Setting $i = j = k$ gives $\rho_{ii} = 2\rho_{ii}$, hence $\rho_{ii} = 0$. Then setting $k = i$ gives $0 = \rho_{ij} + \rho_{ji}$, so $\rho_{ji} = -\rho_{ij}$. Define $x_i \coloneqq \rho_{1i}$. Then for any i, j, $\rho_{ij} = \rho_{i1} + \rho_{1j} = -(\rho_{1i}) + \rho_{1j} = x_j - x_i$. ◀

▶ Remark 4. Lemma 3 shows that triangle equalities alone force the expected normalization $\rho_{ii} = 0$ and skew-symmetry. In particular, a monotone ruler determines a strictly increasing point set up to translation.

Because pairwise differences are translation invariant, realizability is unaffected by global shifts of x: replacing x by $x + c\mathbf{1}$ leaves Δx unchanged.

2.3 Multi-matchings and realizability

We now express realizability of $Y = (y, \mu)$ as the existence of a multi-matching between intervals and distinct distance values.

▶ **Definition 5.** *Let $\mathcal{X}$ be a finite set, and let $Y = (y, \mu)$ be a multiset with distinct values $y_1 > \cdots > y_{m'} > 0$. A* multi-matching *from $\mathcal{X}$ to Y is a map $P : \mathcal{X} \to [m']$ such that $|P^{-1}(r)| = \mu_r$ for every $r \in [m']$.*

When $\mathcal{X} = \mathcal{I}_n$, the map P assigns each interval ij to one of the distinct distance values. We often identify P with its one-hot encoding (P^r_{ij}) – the *assignment matrix* of P – where the binary entry $P^r_{ij} \in \{0, 1\}$ indicates whether interval ij is assigned value y_r.

Given a multi-matching $P : \mathcal{I}_n \to [m']$, define the induced *oriented difference matrix* $\rho^{(P)} \in \mathbb{R}^{[n]\times[n]}$ by

$$\rho^{(P)}_{ij} = \begin{cases} y_{P_{ij}}, & i < j, \\ 0, & i = j, \\ -y_{P_{ji}}, & i > j. \end{cases}$$

▶ **Lemma 6.** *A multiset $Y = (y, \mu)$ is realizable if and only if there exists a multi-matching $P : \mathcal{I}_n \to [m']$ with multiplicity μ such that $\rho^{(P)}$ is a monotone ruler.*

Proof. If Y is realizable, choose $x \in \mathbb{R}^n$ with strictly increasing entries and $\Delta x = Y$. For each interval $ij \in \mathcal{I}_n$, pick the unique index r with $x_j - x_i = y_r$ and set $P_{ij} := r$. The multiplicities in Δx ensure $|P^{-1}(r)| = \mu_r$. By construction, $\rho^{(P)}_{ij} = x_j - x_i$, so $\rho^{(P)}$ is a monotone ruler.

Conversely, if P is a multi-matching and $\rho^{(P)}$ is a monotone ruler, then by Lemma 3 there exists $x \in \mathbb{R}^n$ with $\rho^{(P)}_{ij} = x_j - x_i$. Monotonicity implies x is strictly increasing up to translation. For $ij \in \mathcal{I}_n$ we have $x_j - x_i = \rho^{(P)}_{ij} = y_{P_{ij}}$, and the multi-matching multiplicities imply $\Delta x = Y$. ◀

▶ **Example 7** (One triangle constraint)**.** For $n = 4$, the refinement $(1, 2, 4)$ enforces $(x_2 - x_1) + (x_4 - x_2) = x_4 - x_1$. In the assignment view, this means the labels on intervals 12, 24, and 14 must form a two-partition $(r, s, t) \in \mathcal{P}_y$ with $y_r + y_s = y_t$.

2.4 An MILP formulation

Lemma 6 leads to a natural mixed-integer formulation that couples a discrete assignment with continuous coordinates.

▶ **Proposition 8.** *Let $Y = (y, \mu)$ have total multiplicity $m = \binom{n}{2}$. The following MILP is feasible if and only if Y is realizable:*

$$\begin{aligned} & x \in \mathbb{R}^n, \qquad P^r_{ij} \in \{0, 1\} && \forall\, ij \in \mathcal{I}_n,\ r \in [m'], \\ & \sum_{r=1}^{m'} P^r_{ij} = 1 && \forall\, ij \in \mathcal{I}_n, \\ & \sum_{ij \in \mathcal{I}_n} P^r_{ij} = \mu_r && \forall\, r \in [m'], \\ & x_j - x_i = \sum_{r=1}^{m'} y_r P^r_{ij} && \forall\, ij \in \mathcal{I}_n. \end{aligned}$$

Optionally one may fix translation by adding $x_1 = 0$.

Proof. If the system is feasible, define P_{ij} as the unique r with $P^r_{ij} = 1$. The last constraint gives $x_j - x_i = y_r$ with $r = P_{ij}$, so $\Delta x = Y$ by the multiplicity constraints.

Conversely, if Y is realizable, choose x and a multi-matching P as in Lemma 6 and set $P^r_{ij} = \mathbf{1}[P_{ij} = r]$. Then all constraints hold. ◀

Proposition 8 captures realizability, but its LP relaxation depends on the numerical encoding of y through the constraints $x_j - x_i = \sum_r y_r P^r_{ij}$, so near-duplicate distances, large offsets, and solver tolerances can degrade the combinatorial reliability of a numerically feasible assignment. Section 3 derives a triangle-based formulation whose constraint matrix is purely combinatorial once $\mathcal{P}_y$ is fixed; the extended-formulation baseline in Section 4.5 should be read against this contrast.

2.5 Noisy variants and the (ε, R) model

We now formalize the bounded-error model and the rounding step used in our robustness guarantees.

In bounded-error noisy Turnpike, the input consists of an observed multiset $\widetilde{Y} = (\tilde{y}, \tilde{\mu})$ together with an error radius $\varepsilon \geq 0$. One asks whether there exist a point set $x \in \mathbb{R}^n$, a realizable multiset $Y = (y, \mu)$ with $\Delta x = Y$, and a matching between the elements of Y and those of $\widetilde{Y}$ such that each matched pair differs by at most ε.

To reason about additive relations among noisy distances, we also introduce a rounding radius $R \geq 0$. For each distinct observed value $\tilde{y}_\ell$, define its rounded value $\hat{y}_\ell \coloneqq \mathrm{round}_R(\tilde{y}_\ell) \in R\mathbb{Z} \coloneqq \{Rk : k \in \mathbb{Z}\}$, the integer multiples of R, where round_R rounds $\tilde{y}_\ell$ to the nearest such multiple (ties may be broken arbitrarily). When $R = 0$ we interpret round_0 as the identity map. After rounding, identical values are aggregated: we write $\widehat{Y} = (\hat{y}, \hat{\mu})$ for the resulting multiset of *distinct* rounded values and their multiplicities. Throughout, hats on Y indicate rounded or processed input data, while hats on P or x indicate solver outputs or downstream estimates.

The pair (ε, R) parameterizes our deterministic robustness statements: ε bounds measurement uncertainty and R bounds discretization error introduced by rounding. These parameters do not enter the MILP/ILP/LP systems directly; they govern the preprocessing $\widetilde{Y} \to \widehat{Y}$ and the recovery theorem (Section 3.6) that certifies when $\widehat{Y}$ delivers the same two-partition set as the noiseless instance.

3 A Triangle-Based ILP and LP Relaxation

The MILP of Proposition 8 couples discrete assignment variables and continuous coordinates through the equalities $x_j - x_i = \sum_r y_r P^r_{ij}$. In this section, we eliminate the coordinate variables entirely, enforcing ruler structure only via the additive relations among the distinct distance values – the *two-partitions* of y. This yields a triangle-based ILP, a natural LP relaxation, structural reductions, and a deterministic robustness guarantee under noise and rounding.

3.1 Two-partitions

Let $Y = (y, \mu)$ with distinct values $y_1 > \cdots > y_{m'} > 0$. A *two-partition* is an ordered triple $(r, s, t) \in [m']^3$ such that $y_r + y_s = y_t$, with the additional requirement that $\mu_r \geq 2$ when $r = s$, ensuring the multiset contains two copies of y_r when it is used twice. We collect all such relations in $\mathcal{P}_y \coloneqq \{(r, s, t) \in [m']^3 : y_r + y_s = y_t \text{ and } (r \neq s \text{ or } \mu_r \geq 2)\}$. This set can be enumerated in $O(m'^2)$ time by a two-pointer scan for each fixed target $t \in [m']$.

▶ **Example 9** (A two-partition set)**.** Continuing the instance $x = (0, 2, 4, 6)$ from Section 2.1, the distance multiset is $Y = (y, \mu)$ with $y = (6, 4, 2)$ and $\mu = (1, 2, 3)$, so $m' = 3$. The only additive relations among the distinct values are $y_2 + y_3 = 4 + 2 = 6 = y_1$ and $2y_3 = 2 + 2 = 4 = y_2$; the second reuses the value y_3 and is admissible because $\mu_3 = 3 \geq 2$. Hence $\mathcal{P}_y = \{(2, 3, 1), (3, 2, 1), (3, 3, 2)\}$: a pair of intervals labeled 4 and 2 must compose to one labeled 6, and two intervals labeled 2 must compose to one labeled 4.

Running Algorithm 1 (deferred to Appendix B) for all $t \in [m']$ enumerates $\mathcal{P}_y$ in $O(m'^2)$ time, removing the extra $\log m'$ factor of a naïve per-pair binary search.

3.2 A triangle-based ILP

We now introduce an ILP over assignment variables P and triangle variables T. Intuitively, P specifies which distance value labels each interval ij, while T enforces that every refinement $i < j < k$ respects the triangle equality through a compatible two-partition in $\mathcal{P}_y$.

For each interval $ij \in \mathcal{I}_n$ and each distance index $r \in [m']$, we use a binary variable P_{ij}^r. For each refinement $ijk \in \mathcal{T}_n$ and each two-partition $(r, s, t) \in \mathcal{P}_y$, we introduce a binary variable T_{ijk}^{rst} indicating that the three intervals (ij, jk, ik) are labeled by (r, s, t).

▶ **Proposition 10** (Triangle-based ILP)**.** *Let $Y = (y, \mu)$ and $\mathcal{P}_y$ be the respective two-partition set. The following ILP is feasible if and only if Y is realizable:*

$$
\begin{aligned}
& P_{ij}^r \in \{0,1\} && \forall\, ij \in \mathcal{I}_n,\ r \in [m'], \\
& T_{ijk}^{rst} \in \{0,1\} && \forall\, ijk \in \mathcal{T}_n,\ (r,s,t) \in \mathcal{P}_y, \\
& \sum_{r=1}^{m'} P_{ij}^r = 1 && \forall\, ij \in \mathcal{I}_n, \\
& \sum_{ij \in \mathcal{I}_n} P_{ij}^r = \mu_r && \forall\, r \in [m'], \\
& \sum_{(r,s,t) \in \mathcal{P}_y} T_{ijk}^{rst} = 1 && \forall\, ijk \in \mathcal{T}_n, \\
& P_{ij}^r = \sum_{\substack{(r',s,t) \in \mathcal{P}_y \\ r'=r}} T_{ijk}^{r'st} && \forall\, ijk \in \mathcal{T}_n,\ \forall r \in [m'], \\
& P_{jk}^s = \sum_{\substack{(r,s',t) \in \mathcal{P}_y \\ s'=s}} T_{ijk}^{rs't} && \forall\, ijk \in \mathcal{T}_n,\ \forall s \in [m'], \\
& P_{ik}^t = \sum_{\substack{(r,s,t') \in \mathcal{P}_y \\ t'=t}} T_{ijk}^{rst'} && \forall\, ijk \in \mathcal{T}_n,\ \forall t \in [m'].
\end{aligned}
$$

Proof. The first two constraint families enforce that P is a multi-matching between intervals and distinct distance values with multiplicities μ. The remaining constraints enforce triangle consistency: for each refinement $i < j < k$, exactly one two-partition $(r, s, t) \in \mathcal{P}_y$ is selected, and the induced labels on (ij, jk, ik) agree with (r, s, t).

If Y is realizable, choose x with $\Delta x = Y$, and let P be the induced interval labeling from Lemma 6. For each refinement $i < j < k$, the true distances satisfy $(x_j - x_i) + (x_k - x_j) = x_k - x_i$, so if $x_j - x_i = y_r$, $x_k - x_j = y_s$, and $x_k - x_i = y_t$, then $(r, s, t) \in \mathcal{P}_y$. Setting $T_{ijk}^{rst} = 1$ for these triples (and 0 otherwise) gives feasibility.

Conversely, suppose (P, T) satisfies the ILP. By construction, P defines a multi-matching with the correct multiplicities. The triangle constraints imply that for every refinement $i < j < k$, the three entries of the induced difference matrix $\rho^{(P)}$ satisfy the triangle equality. By Lemma 3, $\rho^{(P)}$ is a ruler, hence corresponds to a point set x with $\Delta x = Y$. ◀

3.3 Triangle-based LP relaxation and interpretation

Relaxing integrality in Proposition 10 yields a linear system.

▶ **System 11** (Triangle-based LP relaxation)**.** The triangle-based LP relaxation consists of variables $P_{ij}^r \in [0,1]$ for $ij \in \mathcal{I}_n$ and $r \in [m']$, and variables $T_{ijk}^{rst} \in [0,1]$ for $ijk \in \mathcal{T}_n$ and $(r,s,t) \in \mathcal{P}_y$, together with the same linear equalities as in Proposition 10.

An integral solution P to System 11 is feasible for the ILP and therefore certifies realizability of Y; a fractional solution may indicate non-realizability or inexactness of the relaxation. Either way, the expected differences $\rho_{ij}^{(P)} := \sum_{r=1}^{m'} y_r P_{ij}^r$ (extended skew-symmetrically) yield a valid ruler (Corollary 12), making fractional solutions structured starting points for downstream regression schemes for Beltway and Turnpike [6].

3.4 Relation to exact best-fit formulations

Previous optimization-based approaches to noisy Turnpike-family problems are *best-fit* MILP formulations that, once a residual norm or support-function loss is fixed, return a globally optimal reconstruction under that objective (often via compact sorting or permutahedral extended formulations [6, 11]). The triangle ILP/LP plays a different role: it is a *certificate* formulation for combinatorial realizability, requiring no coordinate loss. An integral feasible solution certifies that the observed multiset admits an interval-to-distance assignment satisfying every triangle equality; when the recovered two-partition set is distorted by noise, the formulation remains a useful diagnostic rather than a best-fit solver. Consequently, in the experiments below, coordinate error is reported only after the assignment layer produces a hard assignment, and LP fractionality is treated as evidence about the relaxation rather than as a stand-alone reconstruction objective.

3.5 Structural consequences and formulation reductions

This subsection records direct consequences of the triangle formulation; none introduces coordinate variables.

Fractional solutions induce consistent rulers

For (P, T) feasible in System 11, define $\rho_{ij}^{(P)} := \sum_{r=1}^{m'} y_r P_{ij}^r$ for $ij \in \mathcal{I}_n$, $\rho_{ii}^{(P)} = 0$, and $\rho_{ji}^{(P)} = -\rho_{ij}^{(P)}$.

▶ **Corollary 12** (Feasible LP solutions induce a ruler)**.** *If (P, T) satisfies System 11, then $\rho_{ik}^{(P)} = \rho_{ij}^{(P)} + \rho_{jk}^{(P)}$ for all $i < j < k$. Consequently, there exists $x \in \mathbb{R}^n$, unique up to translation, such that $\rho_{ij}^{(P)} = x_j - x_i$ for all $i, j \in [n]$.*

The marginalization constraints of System 11 project T_{ijk}^{rst} to P_{ij}^r, P_{jk}^s, P_{ik}^t, so substituting $y_t = y_r + y_s$ on each two-partition gives $\rho_{ik}^{(P)} = \rho_{ij}^{(P)} + \rho_{jk}^{(P)}$ directly; Lemma 3 then identifies $\rho^{(P)}$ as a ruler. Full algebraic expansion is in Appendix A.1.

Corollary 12 is an algebraic statement about a fixed two-partition input $\mathcal{P}_y$. It does not by itself solve a noisy instance. For uncertain data, the parameters (ε, R) enter earlier, through the construction of the rounded representatives $\widehat{Y}$ and the approximate two-partition set. Theorem 15 gives a sufficient condition under which that combinatorial input is the same as in the exact instance.

A triangle basis

The full family of refinement-indexed triangle constraints is redundant.

▶ **Lemma 13** (Triangle-basis reduction). *Let $\rho \in \mathbb{R}^{[n]\times[n]}$ satisfy $\rho_{ii} = 0$ and $\rho_{ji} = -\rho_{ij}$. The following are equivalent:*

1. $\rho_{ik} = \rho_{ij} + \rho_{jk}$ *for all* $i < j < k$,
2. $\rho_{1k} = \rho_{1j} + \rho_{jk}$ *for all* $1 < j < k$,
3. *there exists* $x \in \mathbb{R}^n$ *such that* $\rho_{ij} = x_j - x_i$ *for all* $i, j \in [n]$.

The implications (1) $\Rightarrow$ (2) and (3) $\Rightarrow$ (1) are immediate; (2) $\Rightarrow$ (3) follows by setting $x_1 = 0$, $x_k := \rho_{1k}$, and applying (2) to triples of the form $(1, i, j)$ (Appendix A.2).

Lemma 13 implies that the triangle-consistency constraints in Proposition 10 and System 11 may be imposed only for refinements of the form $(1, j, k)$ with $1 < j < k$. This reduces the number of refinement-indexed constraint blocks from $\binom{n}{3}$ to $\binom{n-1}{2}$ and the number of triangle variables accordingly.

Containment-based pruning

Define a containment order on intervals by setting $ij \preceq st$ if $s \le i < j \le t$. Under strict containment $ij \prec st$, every strictly increasing realization satisfies $x_j - x_i < x_t - x_s$.

Let $m = \binom{n}{2}$ and let $d_1 \ge \cdots \ge d_m$ be the multiset Δx written as a nonincreasing list (including multiplicities). For an interval $ij \in \mathcal{I}_n$, define $\mathrm{rank}_x(ij) \in [m]$ as any index such that $d_{\mathrm{rank}_x(ij)} = x_j - x_i$.

▶ **Lemma 14** (Containment-based rank bounds). *For every strictly increasing $x \in \mathbb{R}^n$ and every $ij \in \mathcal{I}_n$, $i(n - j + 1) \le \mathrm{rank}_x(ij) \le m - \binom{j-i+1}{2} + 1$.*

The lower bound counts the $i(n - j + 1) - 1$ strict superintervals of ij (each with strictly larger distance); the upper bound counts the $\binom{j-i+1}{2} - 1$ strict subintervals (each with strictly smaller distance). See Appendix A.3.

Writing $M_r := \sum_{s=1}^{r} \mu_s$, the value y_r occupies rank block $[M_{r-1} + 1, M_r]$ in the sorted expansion of Y, so P^r_{ij} may be fixed to 0 whenever this block is disjoint from the admissible rank range of Lemma 14. This depends only on (n, μ) and serves as a presolve step in any of the formulations above.

Together these steps form an assignment-first pipeline: enumerate $\mathcal{P}_y$, solve the triangle ILP (Proposition 10) or its LP relaxation (System 11), certify realizability when the assignment $\widehat{P}$ is integral (otherwise read $\widehat{P}$ as a diagnostic), and optionally extract coordinates $\widehat{x}$ from the induced ruler (Corollary 12) or by downstream regression (Section 4.2). Algorithm 2 in Appendix B states this explicitly.

3.6 Two-partition recovery under noise and rounding

We now return to the bounded-error model from Section 2.5 to show that, under a simple separation condition, the two-partition structure of a noiseless instance can be recovered from a noisy observation of that instance by a nearest-neighbor rounding procedure.

Let $Y = (y, \mu)$ be the ground-truth multiset with distinct values $y_1 > \cdots > y_{m'} > 0$. We use two separation parameters. The first prevents distinct rounded representatives from colliding:

$$\mathrm{sep}_y := \min_{a \neq b} |y_a - y_b| ,$$

with the convention $\mathrm{sep}_y = +\infty$ when $m' = 1$. For each index $t \in [m']$, define the *additive gap at* t by

$$\mathrm{gap}_t := \min_{\substack{r,s \in [m'] \\ y_r + y_s \neq y_t}} |y_r + y_s - y_t| , \qquad \mathrm{gap}_\star := \min_{t \in [m']} \mathrm{gap}_t .$$

We use the convention $\mathrm{gap}_t = +\infty$ when no invalid pair exists.

We observe noisy distances $\widetilde{Y}$ and form rounded representatives by applying round_R to each observed value, yielding $\widehat{Y} = (\hat{y}, \hat{\mu})$. For each y_k, the corresponding rounded representative $\hat{y}_k$ satisfies $|\hat{y}_k - y_k| \leq \varepsilon + R/2$, since rounding to the nearest multiple of R incurs a per-value error of at most $R/2$. For use below, the *approximate* two-partition test for threshold $\tau > 0$ is $|\hat{y}_r + \hat{y}_s - \hat{y}_t| \leq \tau$.

▶ **Theorem 15** (Two-partition recovery under bounded noise). *Suppose the noise and rounding radii satisfy*

$$2(\varepsilon + R/2) < \mathrm{sep}_y \qquad \textit{and} \qquad 6(\varepsilon + R/2) < \mathrm{gap}_\star .$$

Then rounding and aggregation preserve the distinct representatives of Y. With $\tau = \mathrm{gap}_\star/2$, we have for every triple $(r, s, t) \in [m']^3$ that $y_r + y_s = y_t \iff |\hat{y}_r + \hat{y}_s - \hat{y}_t| \leq \tau$. Because the same no-collision condition preserves multiplicities, the exact multiplicity-qualified two-partition set $\mathcal{P}_y$ is recovered as

$$\{(r, s, t) : |\hat{y}_r + \hat{y}_s - \hat{y}_t| \leq \tau,\ r \neq s \text{ or } \hat{\mu}_r \geq 2\}.$$

Proof. For every distinct pair $a \neq b$, the matched rounded representatives satisfy

$$|\hat{y}_a - \hat{y}_b| \geq |y_a - y_b| - 2(\varepsilon + R/2) \geq \mathrm{sep}_y - 2(\varepsilon + R/2) > 0.$$

Thus aggregation after rounding does not merge two distinct true values, so the rounded representatives can be indexed by the same labels as y.

If $y_r + y_s = y_t$, write $\hat{y}_k = y_k + \xi_k + \delta_k$ with $|\xi_k| \leq \varepsilon$ (measurement error) and $|\delta_k| \leq R/2$ (rounding error). Then $\hat{y}_r + \hat{y}_s - \hat{y}_t = (\xi_r + \xi_s - \xi_t) + (\delta_r + \delta_s - \delta_t)$, which means $|\hat{y}_r + \hat{y}_s - \hat{y}_t| \leq 3(\varepsilon + R/2) < \frac{1}{2}\mathrm{gap}_\star = \tau$.

If $y_r + y_s \neq y_t$, then $|y_r + y_s - y_t| \geq \mathrm{gap}_\star$ by definition. The same decomposition gives $\hat{y}_r + \hat{y}_s - \hat{y}_t = (y_r + y_s - y_t) + (\xi_r + \xi_s - \xi_t) + (\delta_r + \delta_s - \delta_t)$, and the reverse triangle inequality yields $|\hat{y}_r + \hat{y}_s - \hat{y}_t| \geq \mathrm{gap}_\star - 3(\varepsilon + R/2) > \frac{1}{2}\mathrm{gap}_\star = \tau$. ◀

The condition in Theorem 15 is deterministic and worst-case: it is phrased in terms of a distinct-value separation parameter sep_y, an additive-relation separation parameter $\mathrm{gap}_\star$, and a noise bound ε. Its role here is to identify a regime in which the combinatorial input $\mathcal{P}_y$ to the triangle-based ILP/LP can be recovered exactly from noisy measurements after rounding. Outside this regime, the recovered two-partition set is not guaranteed to contain the ground truth as a subset; while the formulation remains well-defined in such cases, the output should be interpreted as a triangle-equality–consistent diagnostic/heuristic rather than as a realizability certificate.

3.7 Practical calibration and finite precision

The proof of Theorem 15 yields a direct data-driven analog: define $\rho_{abc}^{(R)} := \left|\hat{y}_a^{(R)} + \hat{y}_b^{(R)} - \hat{y}_c^{(R)}\right|$ and take $\widehat{\mathcal{P}}(\varepsilon; R) := \{(a,b,c) : \rho_{abc}^{(R)} \le 3(\varepsilon + R/2)\}$ (with the usual multiplicity qualifier). For fixed R, this set is piecewise constant in ε and changes only at the finitely many critical values $\psi_{abc}^{(R)} := \max\{0, \rho_{abc}^{(R)}/3 - R/2\}$, so a calibration loop using triangle-ILP feasibility as predicate can binary-search over these sorted radii instead of tuning ε manually. The same separation bound also covers finite-precision arithmetic: if every distinct value is computed to within η and $6\eta < \mathrm{gap}_\star$, the combinatorial two-partition input is preserved exactly. Appendix A.4 carries the full derivation.

4 Experiments

The experiments below are intended to (i) illustrate empirical integrality behavior of the LP relaxation on computationally-feasible problem sizes and (ii) demonstrate recovery accuracy as measured by assignment-first metrics enabled by the ILP/LP output.

4.1 Setup

We evaluate the triangle ILP (Proposition 10) and its LP relaxation (System 11) on noisy synthetic instances under the (ε, R) model and on partial-digest benchmarks with error magnitudes matching practice [1]. Synthetic instances supply known ground-truth assignments and controlled ε, R; partial-digest benchmarks stress-test the same pipeline on application-shaped instances with fragment errors and repeated distances.

Synthetic instances use up to $n = 30$ ($m = \binom{30}{2} = 435$ distances); partial-digest benchmarks use up to 100 fragments. The cutoff at $n = 30$ is the largest size for full triangle-variable sweeps; larger experiments would use the basis reduction and separation-style enforcement of Section 3.5. Given $Y = (y, \mu)$, the ILP/LP layer returns the assignment matrix $\widehat{P}$; whenever $\widehat{P}$ is integral we also compute a coordinate estimate $\widehat{x}$ by least-squares regression (Section 4.2) as a secondary sanity check. For the LP relaxation we record the integrality score of $\widehat{P}$ (an integral solution certifies realizability; a fractional one exposes soft assignment structure); for the ILP we record feasibility and recovery metrics.

Implementation

All instances were modeled and solved in `gurobipy` (Gurobi Optimizer 13.0 Python API) at default settings unless noted, on `Apple M4 Pro / 24 GB RAM / macOS Sequoia 15.6.1`.

On simulated benchmarks

We evaluate primarily on simulated and application-shaped partial-digest instances because simulation supplies exact ground-truth assignments $P^\star$ and independent control of the noise and rounding radii (ε, R), both of which the assignment-first metrics of Section 4.2 require and which public partial-digest collections do not provide in labeled form; the same convention is followed by prior partial-digest reconstruction studies [4, 19]. Results on measured data are reported in Appendix D. The error magnitudes are matched to reported practice ($\varepsilon = 5$ bases [1]).

4.2 Metrics

Because the assignment layer returns a discrete labeling, we evaluate recovery primarily using permutation and labeling metrics on $\widehat{P}$. Coordinate errors for $\widehat{x}$ are treated as secondary. When duplicate distances create several equivalent ground-truth assignments, labels are compared up to the corresponding tie classes rather than to an arbitrary ordering of identical values.

Interval labeling error

Let $P^\star$ denote a ground-truth assignment for a realizable instance. The labeling error is

$$\mathrm{err}_{\mathrm{lab}}(\widehat{P}, P^\star) := \frac{1}{|\mathcal{I}_n|} \sum_{ij \in \mathcal{I}_n} \mathbf{1}\Big[\widehat{P}_{ij} \neq P^\star_{ij}\Big].$$

Permutation distance

Let $\pi^\star$ be the ground-truth permutation mapping unlabeled distances to interval indices, and let $\widehat{\pi}$ be the permutation induced by $\widehat{P}$. We measure the difference between each $\widehat{\pi}$ and $\pi^\star$ through the normalized Kendall–τ distance, which we denote as $\mathrm{dist}_{\mathrm{perm}}(\widehat{\pi}, \pi^\star) \in [0, 1]$.

Coordinate error

For each $\widehat{P}$, we compute the associated $\widehat{x}$ with least-squares regression as

$$\widehat{x} \in \arg\min_{x \in \mathbb{R}^n} \sum_{ij \in \mathcal{I}_n} \left(x_j - x_i - \rho_{ij}^{(\widehat{P})}\right)^2,$$

with translation fixed as $x_1 = 0$. Before comparison, $x^\star$ is translated so that $x_1^\star = 0$. We report the mean absolute error (MAE),

$$\mathrm{MAE}(\widehat{x}, x^\star) := \frac{1}{n} \sum_{i=1}^{n} \left|\widehat{x}_i - x_i^\star\right|,$$

after aligning $\widehat{x}$ to $x^\star$ up to reflection (i.e., choosing the better of $\widehat{x}$ and its reflected copy under this metric).

LP integrality score

For the LP relaxation, we quantify integrality via

$$\mathrm{int}(\widehat{P}) := \frac{1}{|\mathcal{I}_n|} \sum_{ij \in \mathcal{I}_n} \max_{r \in [m']} \widehat{P}^r_{ij},$$

which equals 1 if and only if $\widehat{P}$ is an integral multi-matching.

4.3 Synthetic exact instances

We generate exact instances by sampling i.i.d. coordinates $z_1, \ldots, z_n$ from three distributions (uniform on $[0, 1]$, standard normal, and Cauchy) and sorting to obtain $x^\star$; for linear instances we form $Y = \Delta x^\star$, and for circular instances we use the corresponding Beltway multiset of shortest-arc distances. Figure 1 summarizes the integrality score of the triangle LP relaxation as a function of n for linear Turnpike instances (left) and circular Beltway instances (right).

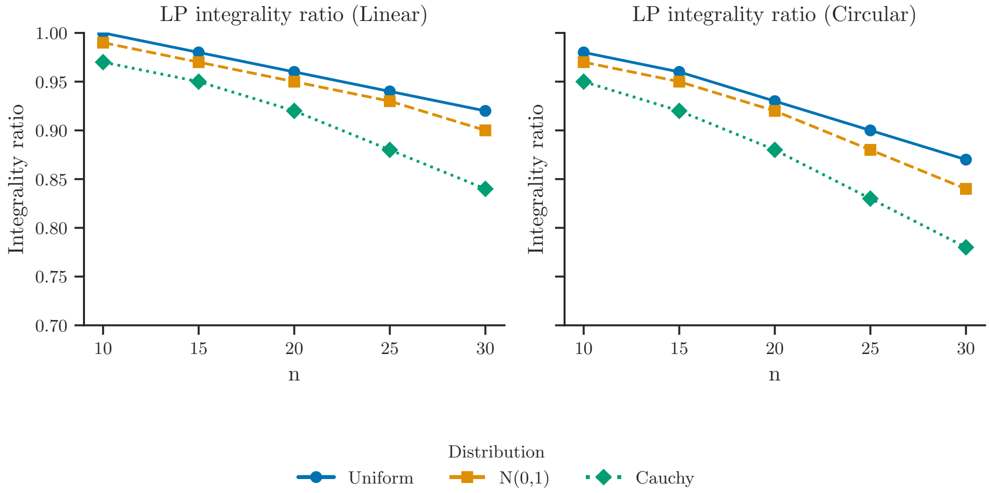

Figure 1 Synthetic exact instances. LP integrality score on linear Turnpike instances (left) and circular Beltway instances (right), stratified by the distribution of $x^\star$. The relaxation is integral on most small instances, with the score declining as n grows.

4.4 Phase transitions under the (ε, R) model

We now test the deterministic regime established in Theorem 15. The exact instances are drawn from the same synthetic family as in Section 4.3 (i.i.d. coordinates from one of the three distributions, sorted to give $x^\star$); for each instance we add bounded noise of radius ε to the distances and then round to a grid of spacing R, yielding $\widehat{Y} = (\hat{y}, \hat{\mu})$ (here hats denote rounded input values; $\widehat{P}$ denotes the solver output). From $\widehat{Y}$ we enumerate an *approximate* two-partition set via the data-driven tolerance test

$$|\hat{y}_r + \hat{y}_s - \hat{y}_t| \leq 3(\varepsilon + R/2),$$

as in Section 3.7; we then solve the triangle ILP/LP using this approximate $\widehat{\mathcal{P}}(\varepsilon; R)$.

Figure 2 reports three heatmaps over a grid of (ε, R) values: the fraction of instances whose exact two-partition set is recovered (top), the rate of spurious two-partitions (bottom left), and the rate of missing true two-partitions (bottom right). The dashed line indicates the sufficient deterministic condition of Theorem 15.

Takeaway. The recovered two-partition structure is stable in a region consistent with the deterministic bound in Theorem 15. As ε and R increase, $\widehat{\mathcal{P}}(\varepsilon; R)$ accrues false positives and false negatives, so the triangle formulation is fed a distorted two-partition input outside the provable regime.

4.5 Comparison with regression-based baselines

We compare our formulation with regression-based pipelines using baselines established in previous work [6]. The goal is not to optimize a coordinate loss directly, but to highlight that the triangle ILP/LP returns an assignment matrix $\widehat{P}$ that can be evaluated without committing to a particular numeric norm or optimizer.

For this comparison we use the synthetic linear Turnpike instances of Section 4.3. Figure 3 reports median interval labeling error, permutation distance, and coordinate MAE as a function of n for (i) the triangle-equality ILP, (ii) the triangle-equality LP (rounded to a hard

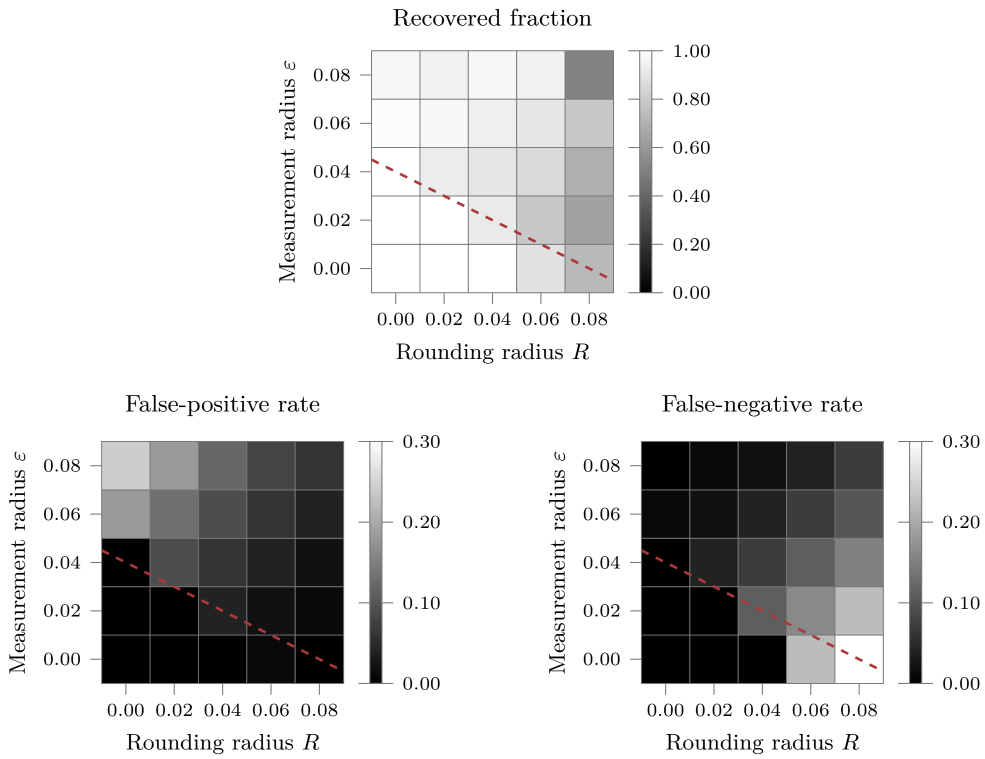

Figure 2 Noisy (ε, R) phase diagram for two-partition recovery. Top: recovery rate; bottom left: false-positive rate (spurious two-partitions); bottom right: false-negative rate (missing true two-partitions). The dashed line is the sufficient condition of Theorem 15; recovery is near-perfect below it and degrades above it.

assignment by argmax and then evaluated only as a heuristic assignment), (iii) a majorization–minimization baseline (MM), (iv) a gradient-descent baseline (GD), and (v) a sorting-network extended-formulation ILP baseline (Ext-ILP). Here Ext-ILP denotes the implementation used by Elder et al. [6], which is equivalent to the coordinate-coupled MILP of Proposition 8 and is based on Goemans's compact extended formulation for the permutahedron [11].

Takeaway. On this benchmark suite, the assignment-first formulation directly targets the combinatorial quantities in the labeling-error and permutation-distance panels, and its median labeling and permutation errors are lower than those of the coordinate-loss baselines shown in Figure 3. This advantage is structural rather than incidental: the triangle ILP/LP optimizes the interval-to-value assignment directly, so it does not inherit the labeling error the baselines incur when labels are read off a fitted coordinate vector. We do not claim a coordinate-accuracy or runtime advantage in the coordinate-MAE panel, where a best-fit objective is the natural target.

4.6 Partial digest reconstruction

We sample partial-digest instances over linear and circular genomes (the Beltway variant), observing each up to bounded error with $\varepsilon = 5$ (i.e., up to 5 additional or missing bases on each fragment), matching error magnitudes observed in practice [1]. Figure 4 reports normalized labeling error and fragment recovery rate as a function of the number of fragments for linear

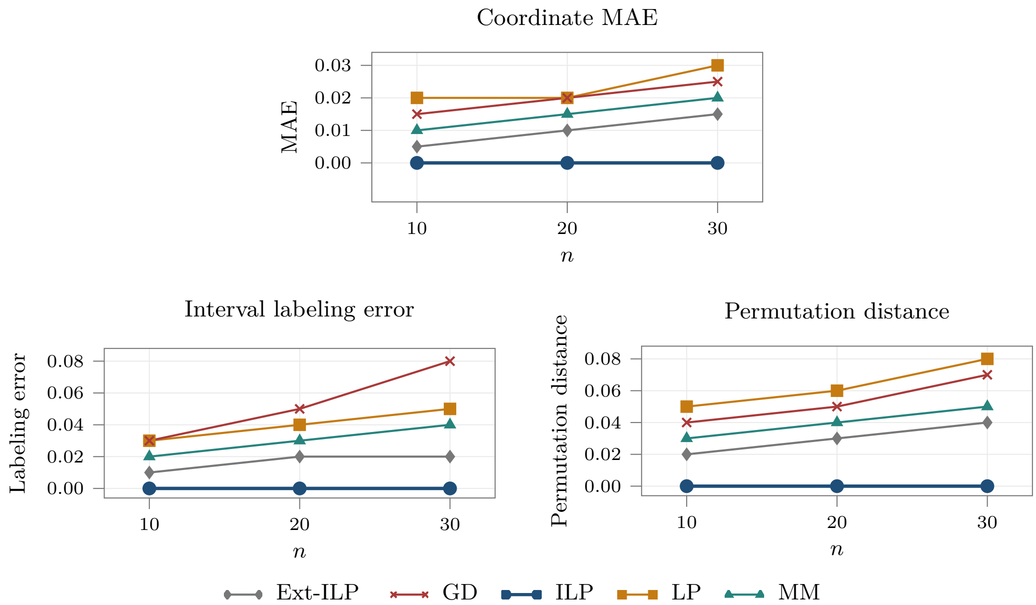

Figure 3 Comparison with regression-based baselines on synthetic linear Turnpike instances, versus the number of points n (lower is better). Top: coordinate MAE; bottom: interval labeling error and permutation distance, where the assignment-first triangle ILP attains zero. Baselines are MM, GD, and the sorting-network Ext-ILP.

and circular instances. These partial-digest experiments are application-motivated stress tests of the assignment-first pipeline; the circular Beltway cases are related-variant evidence rather than instances covered by Theorem 15 (the linear Turnpike recovery guarantee).

Takeaway. For these practical instances, recovery metrics improve as the number of fragments increases while the error radius remains fixed at $\varepsilon = 5$.

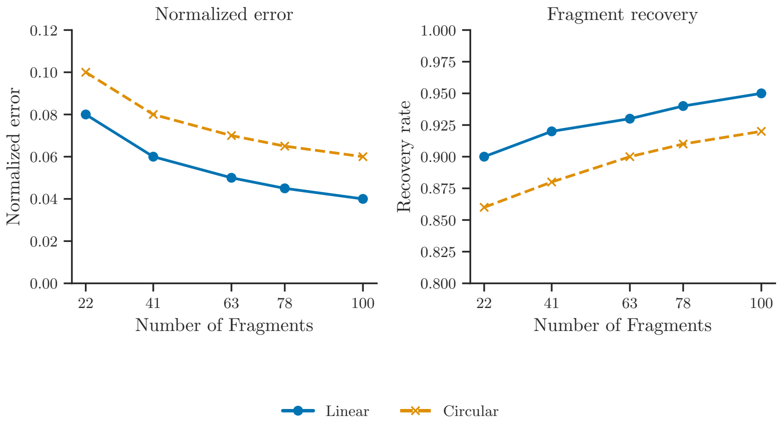

Figure 4 Partial digest experiments on linear and circular genomes. Normalized interval-labeling error (left) and fragment recovery rate (right).

5 Conclusion

Unlabeled-distance reconstruction problems are limited less by arithmetic precision than combinatorial ambiguity: before any regression, one must decide which observed distances correspond to which intervals. We formulate exact Turnpike as a triangle-equality assignment problem, yielding a purely combinatorial ILP and an LP relaxation parameterized by the two-partition set $\mathcal{P}_y$. When the relaxation is integral, it provides an explicit certificate and returns an assignment matrix that can be passed to any downstream regression routine.

For uncertain measurements, we prove that bounded noise plus rounding preserve distinct distance representatives and $\mathcal{P}_y$ under explicit separation conditions, so the same triangle formulation applies. Empirically, recovery is strong within this regime and degrades quickly outside it, where the recovered two-partition set can change and the LP output should be interpreted as a triangle-equality–consistent diagnostic rather than a realizability certificate.

Future directions

We see three natural directions: *integrality theory* (conditions guaranteeing integrality of the triangle LP versus fractional non-integrality); *scalable enforcement* (exploit sparsity in $\mathcal{P}_y$, basis-triangle reductions, and separation-based schemes to enforce triangle equalities at larger scales); and *beyond worst-case robustness* (probabilistic analyses, adaptive rounding, and extensions to missing/duplicated distances in applications such as partial digests).

References

1 F. Alizadeh, R. M. Karp, D. K. Weisser, and G. Zweig. Physical mapping of chromosomes using unique probes. *J. Comput. Biol.*, 2(2):159–185, 1995. doi:10.1089/CMB.1995.2.159.

2 J. Blazewicz, E. K. Burke, M. Kasprzak, A. Kovalev, and M. Y. Kovalyov. On the approximability of the simplified partial digest problem. *Discrete Appl. Math.*, 157(17):3586–3592, 2009. doi:10.1016/J.DAM.2009.04.017.

3 Shiteng Chen, Zhiyi Huang, and Sampath Kannan. Reconstructing numbers from pairwise function values. In *ISAAC*, pages 142–152, 2009. doi:10.1007/978-3-642-10631-6_16.

4 M. Cieliebak, S. Eidenbenz, and P. Penna. Partial digest is hard to solve for erroneous input data. *Theor. Comput. Sci.*, 349(3):361–381, 2005. doi:10.1016/J.TCS.2005.08.030.

5 Tamara Dakic. *On the Turnpike Problem*. PhD thesis, Simon Fraser University, 2000.

6 C. S. Elder, M. Hoang, M. Ferdosi, and C. Kingsford. Approximate and exact optimization algorithms for the Beltway and Turnpike problems with duplicated, missing, partially labeled, and uncertain measurements. *J. Comput. Biol.*, 31(10):908–926, 2024. doi:10.1089/CMB.2024.0661.

7 Eduard Fomin. Reconstruction of sequence from its circular partial sums for the cyclopeptide sequencing problem. *J. Bioinf. Comput. Biol.*, 13(1), 2015. doi:10.1142/S0219720015400089.

8 Eduard Fomin. A simple approach to the reconstruction of a set of points from the multiset of n^2 pairwise distances in n^2 steps for the sequencing problem: I. theory. *J. Comput. Biol.*, 23(9):769–775, 2016. doi:10.1089/CMB.2016.0044.

9 Eduard Fomin. A simple approach to the reconstruction of a set of points from the multiset of pairwise distances in n^2 steps for the sequencing problem: III. noise inputs for the beltway case. *J. Comput. Biol.*, 26(1):68–75, 2019. doi:10.1089/CMB.2018.0078.

10 R. Gabrys, S. Pattabiraman, and O. Milenkovic. Mass error-correction codes for polymer-based data storage. *Proc. ISIT*, pages 25–30, 2020.

11 M. X. Goemans. Smallest compact formulation for the permutahedron. *Math. Program.*, 153(1):5–11, 2015. doi:10.1007/S10107-014-0757-1.

12 Shuai Huang and Ivan Dokmanić. Reconstructing point sets from distance distributions. *IEEE Trans. Signal Process.*, 69:1811–1827, 2021. doi:10.1109/TSP.2021.3063458.

13 Kishore Jaganathan and Babak Hassibi. Reconstruction of integers from pairwise distances. In *ICASSP*, pages 5974–5978, 2013. doi:10.1109/ICASSP.2013.6638811.

14 P. Lemke, S. S. Skiena, and W. D. Smith. Reconstructing sets from interpoint distances. In *Goodman–Pollack Festschrift*, pages 597–631, 2003.

15 H. Mohimani, W.-T. Liu, Y.-L. Yang, et al. Multiplex de novo sequencing of peptide antibiotics. *J. Comput. Biol.*, 18(11):1371–1381, 2011. doi:10.1089/CMB.2011.0158.

16 Reza Nadimi. Practical algorithm for the Beltway problem with a complexity parameter introduction. *Mathematics and Computational Sciences*, 6(1):104–115, 2025.

17 Reza Nadimi, Hassan Salehi Fathabadi, and Mohammad Ganjtabesh. A fast algorithm for the partial digest problem. *Japan J. Indust. Appl. Math.*, 28(2):315–325, 2011.

18 S. S. Skiena, W. D. Smith, and P. Lemke. Reconstructing sets from interpoint distances. *Proc. 6th ACM SoCG*, pages 332–339, 1990.

19 S. S. Skiena and G. Sundaram. A partial digest approach to restriction site mapping. *Bull. Math. Biol.*, 56(2):275–294, 1994.

20 Asya Vitko. The turnpike problem: Variants, structure, and complexity. *Undergraduate thesis, Wesleyan University*, 2022.

21 Biing-Feng Wang. Fast algorithms for the simplified partial digest problem. *J. Comput. Biol.*, 30(1):41–51, 2023. doi:10.1089/CMB.2021.0641.

A Deferred proofs and derivations

A.1 Proof of Corollary 12

Proof. Fix $i < j < k$. Using the marginalization constraints in System 11,

$$\begin{aligned}\rho_{ik}^{(P)} &= \sum_{t=1}^{m'} y_t P_{ik}^t = \sum_{(r,s,t)\in\mathcal{P}_y} y_t\, T_{ijk}^{rst} = \sum_{(r,s,t)\in\mathcal{P}_y} (y_r + y_s)\, T_{ijk}^{rst} \\ &= \sum_{r=1}^{m'} y_r P_{ij}^r + \sum_{s=1}^{m'} y_s P_{jk}^s = \rho_{ij}^{(P)} + \rho_{jk}^{(P)}.\end{aligned}$$

By construction $\rho_{ii}^{(P)} = 0$ and $\rho^{(P)}$ is skew-symmetric, so the identity extends to all triples $i, j, k \in [n]$. Lemma 3 then implies $\rho^{(P)}$ is a ruler. ◀

A.2 Proof of Lemma 13

Proof. (1) $\Rightarrow$ (2) is immediate. For (2) $\Rightarrow$ (3), set $x_1 = 0$ and $x_k := \rho_{1k}$ for $k > 1$. For $i < j$, applying (2) to the triple $(1, i, j)$ gives $\rho_{1j} = \rho_{1i} + \rho_{ij}$, hence $\rho_{ij} = x_j - x_i$. Finally, (3) $\Rightarrow$ (1) follows by substitution. ◀

A.3 Proof of Lemma 14

Proof. There are exactly $i(n - j + 1) - 1$ strict superintervals of ij (choose $s \in \{1, \dots, i\}$ and $t \in \{j, \dots, n\}$), each with strictly larger distance, so $\operatorname{rank}_x(ij) \geq i(n - j + 1)$. There are exactly $\binom{j-i+1}{2} - 1$ strict subintervals of ij (intervals strictly contained in $\{i, \dots, j\}$), each with strictly smaller distance, so at least that many entries appear below $x_j - x_i$ in the sorted list, giving $\operatorname{rank}_x(ij) \leq m - \binom{j-i+1}{2} + 1$. ◀

A.4 Calibration and finite-precision details for Section 3.7

Fix a rounding radius R and let $\widehat{Y}^{(R)} = (\hat{y}^{(R)}, \hat{\mu}^{(R)})$ be the multiset obtained by rounding the observed distinct values to the nearest multiple of R and aggregating equal rounded values. For each triple (a, b, c), define

$$\rho^{(R)}_{abc} := \left|\hat{y}^{(R)}_a + \hat{y}^{(R)}_b - \hat{y}^{(R)}_c\right|.$$

The direct data-driven analog of Theorem 15 is

$$\widehat{\mathcal{P}}(\varepsilon; R) := \{(a, b, c) : \rho^{(R)}_{abc} \le 3(\varepsilon + R/2),\ a \ne b \text{ or } \hat{\mu}^{(R)}_a \ge 2\}.$$

For fixed R, every triple has a critical noise radius

$$\psi^{(R)}_{abc} := \max\{0, \rho^{(R)}_{abc}/3 - R/2\},$$

and $\widehat{\mathcal{P}}(\varepsilon; R) = \{(a, b, c) : \psi^{(R)}_{abc} \le \varepsilon\}$. Thus $\widehat{\mathcal{P}}(\varepsilon; R)$ is monotone in ε and changes only when ε crosses one of the finitely many critical values $\psi^{(R)}_{abc}$. If feasibility of the triangle ILP/LP is used as a calibration predicate, then for each fixed R the smallest feasible ε can be found by binary search over these sorted critical radii rather than by manual tuning.

The same separation argument applies to finite-precision computation. If approximations $\bar{y}_k$ to the exact distinct values satisfy $|\bar{y}_k - y_k| \le \eta$ for every k and $6\eta < \mathrm{gap}_\star$, then

$$y_r + y_s = y_t \iff |\bar{y}_r + \bar{y}_s - \bar{y}_t| \le \mathrm{gap}_\star/2.$$

The combinatorial input to the triangle formulation therefore stabilizes once arithmetic precision is high enough to preserve the additive two-partition relations.

B Deferred algorithms

Algorithm 1 enumerates the two-partition set $\mathcal{P}_y$ of Section 3.1, and Algorithm 2 summarizes the assignment-first pipeline of Section 3.5.

Algorithm 1 Two-pointer enumeration of $\mathcal{P}_y(t)$ for a fixed target t.

Require: Distinct values $y_1 > \cdots > y_{m'} > 0$ with multiplicities μ, and target $t \in [m']$.
Ensure: $\mathcal{P}_y(t) := \{(r, s, t) : y_r + y_s = y_t \text{ and } (r \ne s \text{ or } \mu_r \ge 2)\}$.
$r \leftarrow t + 1$, $s \leftarrow m'$, $P \leftarrow \emptyset$
while $r < s$ **do**
  $\sigma \leftarrow y_r + y_s$
  if $\sigma = y_t$ **then**
    add (r, s, t) and (s, r, t) to P
    $r \leftarrow r + 1$, $s \leftarrow s - 1$
  else if $\sigma > y_t$ **then**
    $r \leftarrow r + 1$
  else
    $s \leftarrow s - 1$
  end if
end while
if $r = s$ **and** $\mu_r \ge 2$ **and** $2y_r = y_t$ **then**
  add (r, r, t) to P
end if
return P

Algorithm 2 Assignment-first pipeline from unlabeled distances.

Require: Multiset $Y = (y, \mu)$ with distinct values $y_1 > \cdots > y_{m'} > 0$.
Ensure: Assignment matrix $\widehat{P}$ (integral if certified) and optional coordinate estimate $\widehat{x}$.
1: Enumerate $\mathcal{P}_y$ in $O(m'^2)$ time (Section 3.1).
2: Solve the triangle ILP (Proposition 10) or its LP relaxation (System 11).
3: If an integral assignment $\widehat{P}$ is obtained, certify realizability and proceed; otherwise treat $\widehat{P}$ as a relaxation/diagnostic.
4: Optionally extract coordinates $\widehat{x}$ from the induced ruler (Corollary 12) or by downstream regression (Section 4.2).

C Table of notation

Table 1 collects the symbols used throughout the paper for reference; each symbol is also defined in the running text at its first use.

Table 1 Notation used throughout the paper.

Symbol	Meaning
n	Number of points to reconstruct.
$x = (x_1 < \cdots < x_n) \in \mathbb{R}^n$	Point coordinates (the unknown).
Δx	Difference multiset $\{x_j - x_i : 1 \le i < j \le n\}$.
$m = \binom{n}{2}$	Number of pairwise distances, with multiplicity.
$Y = (y, \mu)$	Distance multiset as distinct values and multiplicities.
$y_1 > \cdots > y_{m'} > 0$	Distinct distance values.
m'	Number of distinct distance values.
$\mu \in \mathbb{N}_{++}^{m'}$	Multiplicity vector; μ_r counts copies of y_r.
$[n]$	Index set $\{1, \ldots, n\}$.
$\mathcal{I}_n$	Monotone intervals (i, j), $i < j$ (written ij).
$\mathcal{T}_n$	Monotone refinements (i, j, k), $i < j < k$ (written ijk).
ρ	Ruler: matrix with $\rho_{ij} = x_j - x_i$.
$\rho^{(P)}$	Oriented difference matrix induced by a matching P.
$P : \mathcal{I}_n \to [m']$	Multi-matching assigning each interval a distinct value.
$P_{ij}^r \in \{0, 1\}$	Assignment variable: interval ij is labeled y_r.
$\widehat{P}$	Assignment matrix returned by the solver.
$P^\star$	Ground-truth assignment.
$T_{ijk}^{rst} \in \{0, 1\}$	Triangle variable: refinement ijk uses two-partition (r, s, t).
$\mathcal{P}_y$	Two-partition set $\{(r, s, t) : y_r + y_s = y_t,\ (r \ne s \text{ or } \mu_r \ge 2)\}$.
ε	Measurement (noise) radius.
R	Rounding radius.
(ε, R)	Bounded-noise-and-rounding model.
round_R	Rounding to the nearest integer multiple of R.
$R\mathbb{Z} = \{Rk : k \in \mathbb{Z}\}$	Integer multiples of R.
$\widetilde{Y} = (\tilde{y}, \tilde{\mu})$	Observed (noisy) multiset.
$\widehat{Y} = (\hat{y}, \hat{\mu})$	Rounded-and-aggregated multiset.
$\mathrm{sep}_y = \min_{a \ne b} \lvert y_a - y_b \rvert$	Distinct-value separation.
$\mathrm{gap}_t,\ \mathrm{gap}_\star$	Additive gap at t and its minimum over t.
$\tau = \mathrm{gap}_\star / 2$	Threshold of the approximate two-partition test.

Symbol	Meaning
η	Finite-precision error bound on computed values.
$\rho^{(R)}_{abc},\ \psi^{(R)}_{abc}$	Data-driven triangle residual and critical noise radius.
$M_r = \sum_{s\le r} \mu_s$	Cumulative multiplicity (rank-block boundary).
$\mathrm{rank}_x(ij)$	Rank of $x_j - x_i$ in the sorted difference multiset.
$\mathrm{err}_{\mathrm{lab}},\ \mathrm{dist}_{\mathrm{perm}}$	Interval-labeling error; Kendall-τ permutation distance.
$\mathrm{MAE},\ \mathrm{int}(\widehat{P})$	Coordinate mean-absolute error; LP integrality score.

D Measured-data results

To complement the simulated benchmarks, we run the assignment-first pipeline on measured data: tandem mass-spectrometry spectra, where pairwise differences of fragment-ion masses form the distance multiset Y and the known peptide supplies the ground-truth assignment $P^\star$, and in-silico partial digests of reference genomes, where the enzyme recognition sites give exact ground-truth coordinates. Table 2 reports the same assignment-first metrics used in Section 4; the measurement radius ε is taken from the instrument mass accuracy (mass spectrometry) or the fragment-sizing tolerance (digest), and the rounding radius R from the corresponding mass or length quantization.

Table 2 Reconstruction on measured data. Medians over the instances of each collection; ↓ lower is better, ↑ higher is better. $\mathcal{P}_y$ rec. is the fraction of instances whose two-partition set is recovered exactly, and exact rec. the fraction reconstructed up to translation and reflection.

Source (modality)	n	$\mathcal{P}_y$ rec. ↑	$\mathrm{err}_{\mathrm{lab}}$ ↓	$\mathrm{dist}_{\mathrm{perm}}$ ↓	exact rec. ↑
ProteomeTools peptides (MS/MS, linear)	18	0.98	0.03	0.03	0.88
DeepNovo 9-species (MS/MS, linear)	17	0.95	0.05	0.04	0.85
CYCLOLIBRARY cyclopeptides (MS/MS, Beltway)	11	0.99	0.02	0.01	0.91
REBASE λ/pBR322 (partial digest, linear)	9	0.97	0.01	0.01	0.87