Reconstructing Point Sets from Distance Distributions

Shuai Huang ∗ Ivan Dokmanić ∗

April 2018

Abstract

We study the problem of reconstructing the locations $u_i$ of a set of points from their unlabeled pairwise distance measurements. Instead of recovering $u_i$ directly, we represent the point-set by its indicator vector $x \in \{0,1\}^M$ and search for an $x$ that reproduces the observed distance distribution. We show that the integer constraint on $x$ can be further relaxed, and recast the unassigned distance geometry problem into a constrained nonconvex optimization problem. We propose a projected gradient descent algorithm to solve it, and derive conditions under which the proposed method converges to a global optimizer $x^*$ in both noiseless and noisy cases. In addition, we propose several effective initialization strategies that empirically perform well. Compared to conventional greedy build-up approaches that become inoperable in the face of measurement noise, the proposed distance distribution matching approach jointly reconstructs all the sample points and is robust to noise. We substantiate these claims with a number of numerical experiments.

Index terms— unassigned distance geometry problem, turnpike problem, nonconvex optimization, pair distribution function, molecular structure.

1 Introduction

The unassigned distance geometry problem (uDGP) \cite{9,27,28} tries to reconstruct the locations of $N$ samples $U = \{u_i \in \mathbb{R}^D | 1 \leq i \leq N\}$ from their unlabeled/unassigned pairwise distance measurements $Y = \{y_i \in \mathbb{R} | 1 \leq i \leq |Y|\}$, as is illustrated in Figure 1. For each distance measurement $y_i$, no label is assigned to it to tell us which sample pair it belongs to: we have to fill in the blank during the recovery process. uDGP arises in various scenarios where recording the distance labels becomes either impractical or impossible. For example, determining the 3D atomic structure of molecules has been an everlasting topic in fields such as solid state physics, biochemistry, etc., and the measurements we work with are often in the form of unlabeled pairwise atomic distances.

• In the nanostructure problem \cite{8,11,55}, the nanometer-scale compound is studied from its x-ray diffraction pattern. Since the powdered compound lacks the periodic long-range order of a bulk crystal, the obtained powder diffraction pattern is orientationally averaged and the directional information is lost in the process. This ultimately produces a 1D atomic pair distribution function \cite{7} that represents the weighted inter-atomic distance distribution.

∗Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801, U.S.A. Email: huang816@illinois.edu; dokmanic@illinois.edu.
In nuclear magnetic resonance (NMR) spectroscopy \cite{16, 34, 38, 63}, the nuclear overhauser effect \cite{52, 53} only occurs between atoms that are in close proximity, and can thus be exploited to measure the pairwise atomic distances. Due to the lack of labeling information, it is used to simply confirm the refined 3D molecular structures in nuclear magnetic resonance (NMR) spectroscopy \cite{14, 37, 40}.

Additionally, an emerging topic in room acoustics nowadays deals with the recovery of room geometry from echoes picked up by a microphone array \cite{19, 23–26, 33, 44, 48, 57}. The arrival times of the echo impulses can be translated to unlabeled distances between the microphones and “mirror” images of the speaker(s) via the image source model \cite{4, 26}. These unlabeled distance measurements are then used to reconstruct the locations of the speakers and microphones that could reveal the desired room geometry.

1.1 Prior work

Historically, one of the first instances of uDGP was the partial digestion problem in small-scale DNA mapping \cite{12, 22, 32, 49, 56, 64}, also known as the turnpike reconstruction problem in computer science \cite{12, 22, 32, 64} (see Figure 2). Skiena, Smith, and Lemke \cite{41, 61} devised clever exhaustive search methods for the turnpike based on backtracking—building an efficient search tree over permutations so that infeasible branches can hopefully be discarded early in the process. While backtracking is known to require exponential time for certain problem instances, Skiena, Smith, and Lemke \cite{41, 61} observed that the average runtime is favorable \cite{61}. They also proposed based on polynomial factorization \cite{41}. Other exhaustive search strategies have been proposed \cite{1, 2, 13, 21, 51, 58}, however, both polynomial factorization and equality constraint testing are brittle in the presence of noise, even when the solution is a priori unique \cite{15}. Noisy data are known to make the turnpike reconstruction problem NP-hard \cite{17}. Skiena and Sundaram have extended backtracking to handle small noise \cite{62}, however such methods...
cannot perform optimally. As they point out, their approach works well for small problem sizes.

A closely related problem to the turnpike with points arranged on a circle instead of a line is called the beltway problem [61]. It is recently experiencing a revival in de novo cyclic peptide sequencing from mass spectrograms with applications in antibiotic discovery [45–47].

Recent approaches to the uDGP are also clever search methods, primarily bottom-up (globally rigid build-up [10]). They start by combining a few distances to assemble small rigid cores. If the distance list contains many unique distances, then those small cores will participate in larger structures with high likelihood. Finding a core is the hardest part: once it is found augmenting it with new vertices becomes simpler. A notable example of such a strategy is the TRIBOND algorithm [29] which starts by searching for four-vertex cores in 2D. Its main drawback is that even tiny noise makes it unstable to the extent that it gives useless output; its computational complexity also scales unfavorably in 3D. It works well when most distances are unique. An overlapping group of authors earlier proposed a randomized build-up algorithm called LIGA [35, 36]. In LIGA the cores are not required to be perfect: there may be some residual tension in the sense that we may have to bend a few sticks to make them fit. Many such structures are generated and then promoted and demoted in a tournament-like fashion. Further, bad vertices which cause tension are removed stochastically. At its core LIGA is a genetic algorithm so it is challenging to analyze it theoretically. While it does not work well in configurations with many unique distances and comes without performance guaranteed, it successfully reconstructed several known structures ab initio [35].

1.2 Proposed approach

In this paper we take a different perspective on the uDGP by first summarizing the unlabeled distance measurements \( Y \) with the observed distance distribution \( \{P(y) \mid y \in \mathcal{R}\} \):

\[
P(y) := \frac{1}{|Y|} \sum_{i=1}^{|Y|} \delta(y - y_i),
\]

where \( \delta(\cdot) \) is the Dirac delta function. The uDGP then amounts to finding a sample set \( \mathcal{U} \) so that the estimated distance distribution \( Q(y) \) produced by \( \mathcal{U} \) matches the observed distribution \( P(y) \):

\[
\min_{\mathcal{U}} J(P(y), Q(y)),
\]

where \( J(\cdot) \) is some suitable loss function that quantifies the difference between the two distributions. In order to write \( Q(y) \) in a closed form, we first divide the domain space \( \mathbb{R}^D \) into \( M \) unit cells, and represent the \( N \) sample locations in \( \mathcal{U} \) with an indicator vector \( \mathbf{x} \in \{0, 1\}^M \), as is shown in Fig. 3. Each entry \( x_m \) of \( \mathbf{x} \) will correspond to one unit cell in the domain space \( \mathbb{R}^D \). The entry value \( x_m = 0 \) indicates that no sample occupies the \( m \)-th cell, while \( x_m = 1 \) indicates that there is one sample \( u \in \mathcal{U} \) occupying the \( m \)-th cell. As is detailed in Section 2, the integer constraint on \( \mathbf{x} \) can be further relaxed. The uDGP then becomes the following constrained nonconvex optimization problem with respect to \( \mathbf{x} \):

\[
\min_{\mathbf{x}} J(x) = J(P(y), Q(y))
\]

subject to

\[
\|\mathbf{x}\|_1 = N, \quad 0 \leq x_m \leq 1, \quad \forall m \in [M].
\]
Although (1.3) is nonconvex, a good recovery can still be obtained using the proposed projected gradient descent method with proper initialization. In the noiseless recovery case, we can show that the proposed projected gradient descent update is guaranteed to converge linearly to any global optimizer $\mathbf{x}^*$ once it enters a small neighbourhood $\rho(\mathbf{x}^*)$ around $\mathbf{x}^*$. In the noisy recovery case, there is not necessarily such a neighbourhood. However, there are regions next to $\mathbf{x}^*$ that support the convergence to $\mathbf{x}^*$.

2 Distance distribution matching

Previous work on the uDGP treats the multiset $\mathcal{Y}$ as a collection of positive pairwise distances, in this paper we define a more generalized multiset $\mathcal{Y}$ as follows:

- We include the distance from $\mathbf{u}_i$ to itself: $y_{\mathbf{u}_i \rightarrow \mathbf{u}_i} = 0$.
- We differentiate the distance from $\mathbf{u}_i$ to $\mathbf{u}_j$ and the distance from $\mathbf{u}_j$ to $\mathbf{u}_i$: $y_{\mathbf{u}_i \rightarrow \mathbf{u}_j} = y_{\mathbf{u}_j \rightarrow \mathbf{u}_i}$.

The total number of pairwise distances is then $|\mathcal{Y}| = N^2$, where $N = |\mathcal{U}|$ is the number of sample points. Without loss of generality, let $K \in \mathbb{Z}$ denote the largest distance in $\mathcal{Y}$, the sample pair $(\mathbf{u}_1, \mathbf{u}_N)$ corresponding to $y_{\text{max}} = K$ can then be determined (up to certain translation) in the domain space $\mathbb{R}^D$. Using $(\mathbf{u}_1, \mathbf{u}_N)$ as the anchor points, $\mathbb{R}^D$ can be divided into $M$ cells by setting the smallest space unit $e_s = 1$, as is illustrated in Fig. 3.
Let $M$ denote the total number of unit cells in the domain space $\mathbb{R}^D$, the sample locations $\mathcal{U}$ can be represented by an indicator vector $x \in \{0, 1\}^M$. Each entry $x_m$ of $x$ corresponds to one unit cell in the $\mathbb{R}^D$. The entry value $x_m = 0$ indicates that the sample $u$ does not occupy the $m$-th cell $c_m$, while $x_m = 1$ indicates that $u$ occupies $c_m$. In other words, there is a one-to-one mapping between every sample location $u_i \in \mathcal{U}$ and some unit cell $c_m$ in the domain space.

$$u_i \iff c_m. \quad (2.1)$$

Since there are $N$ samples in $\mathcal{U}$, the indicator vector $x$ has a fixed $l_1$ norm $\|x\|_1 = N$. Let $\mathcal{R}$ be the set containing all possible distance values in the domain space $\mathbb{R}^D$, the frequency $G(y)$ of some distance value $y \in \mathcal{R}$ can be written in closed-form with respect to $x$:

$$G(y) = x^T A_y x = Q(y) \cdot \sum_{y \in \mathcal{R}} x^T A_y x, \quad (2.2)$$

where $A_y \in \{0, 1\}^{M \times M}$ is a symmetric matrix containing only $\{0, 1\}$ entry values. Let $y_{c_i \rightarrow c_j}$ denote the distance from the $i$-th cell $c_i$ to the $j$-th cell $c_j$, the $(i, j)$-th entry of $A_y$ is defined as follows:

$$A_{i,j}(y) = \begin{cases} 1, & \text{if } y_{c_i \rightarrow c_j} = y \\ 0, & \text{if } y_{c_i \rightarrow c_j} \neq y. \end{cases} \quad (2.3)$$

We can easily verify that $A = \sum_{y \in \mathcal{R}} A_y$ is a matrix of all 1s. Since $x_m \in \{0, 1\}$ and $\|x\|_1 = N$, we have $\sum_{y \in \mathcal{R}} x^T A_y x = N^2$. (2.2) can be further simplified:

$$G(y) = x^T A_y x = Q(y) \cdot N^2. \quad (2.2 \text{ simplified})$$

Let $x^*$ be an global optimizer that produces the observed frequency measure $F(y)$ and the observed distribution measure $P(y)$:

$$F(y) = x^{*^T} A_y x^* = P(y) \cdot N^2. \quad (2.4)$$

We can see that the frequency measure and distribution measure are equivalent after we introduced the generalized multiset $\mathcal{Y}$.

The uDGP can then be solved by searching for an indicator vector $x$ so that the estimated measures $\{G(y), Q(y)\}$ match the observed measures $\{F(y), P(y)\}$. Specifically, we can minimize a suitable loss function $J(\cdot)$ that quantifies the difference between the estimated and observed distance distribution. From (2.2), (2.4), we can see that the frequency measure and the distribution measure are essentially equivalent. In the following we first present the problem formulation in terms of the distribution measure, then simplify it using the frequency measure.

Various loss functions [3, 6, 18, 20, 31, 42, 43, 50] exist in the literature to quantify the difference between two probability distributions. Here we minimize the following four representative functions to find the solution $\tilde{x}$.

- **Squared Euclidean distance:**

$$\tilde{x} = \arg \min_x \sum_{y \in \mathcal{R}} (Q(y) - P(y))^2$$

$$= \arg \min_x \sum_{y \in \mathcal{R}} \frac{1}{4} (x^T A_y x - F(y))^2. \quad (2.5)$$

[1]Since the largest distance is $K$, the domain space is finite.
• Hellinger distance:

$$\hat{x} = \arg\min_x \sum_{y \in \mathcal{R}} \left( \sqrt{Q(y)} - \sqrt{P(y)} \right)^2$$

$$= \arg\min_x \sum_{y \in \mathcal{R}} -\sqrt{F(y)} \cdot x^T A_y x.$$  \hspace{1cm} (2.6)

• Cross entropy:

$$\hat{x} = \arg\min_x \sum_{y \in \mathcal{R}} -P(y) \log Q(y)$$

$$= \arg\min_x \sum_{y \in \mathcal{R}} -\frac{1}{2} F(y) \log x^T A_y x.$$  \hspace{1cm} (2.7)

The frequency measure $F(y) = x^* A_y x^*$ can be viewed as a quadratic measurement on the true indicator vector $x^*$, and $A_y$ is then the corresponding measurement matrix. (2.5)-(2.7) achieve a global optimizer if and only if the estimated measures $\{G(y), Q(y)\}$ match the observed measures $\{F(y), P(y)\}$. The indicator vector $x$ is characterized by the following constraints:

$$\|x\|_1 = N$$  \hspace{1cm} (2.8)

$$x_m \in \{0, 1\}, \quad \forall m \in [M].$$  \hspace{1cm} (2.9)

We can further relax the constraint (2.9) imposed on entry values of $x$ to:

$$x_m \in [0, 1], \quad \forall m \in [M].$$  \hspace{1cm} (2.10)

**Noiseless recovery:** In the following Lemma 2.1, we show that the relaxed constraint (2.10) does not affect the global optimizers of (1.3).

**Lemma 2.1** Any global optimizer $x$ of the relaxed optimization problem with constraints (2.8),(2.10) has entries that are either 0 or 1 in the absence of noise.

**Proof** Please refer to Appendix A.1.

**Noisy recovery:** In this case the indicator vector $\hat{x}$ recovered from noisy measurements is no longer a 0-1 vector, its entry value $x_m$ could vary between 0 and $N$. However, in practice it is highly unlikely that $x_m$ would reach $N$, and the best performance is usually obtained when we assume $x_m \in [0, r]$, where $r > 0$ is some value around 1. In the following discussion, the entry value $x_m$ is still assumed to vary between 0 and 1, i.e. the relaxed constraint in (2.10). We should note that the analysis still holds for the general case $x_m \in [0, r]$ with minor modifications.

The noisy distance measurements $y$ could come in a continuous or discrete form. For example, in the nanostructure problem the measurement we get corresponds to the pair-distribution function of the inter-atomic distances [7, 54], which can be further processed to get a continuous distance distribution $P(y)$. On the other hand, in order to achieve better recovery performance, the discrete noisy measurement $y$ is further smoothed by convolving it with some pre-specified noise distribution $R(w)$, as is illustrated in Fig. 4.
The discrete noisy measurements are convolved with the noise distribution $R(w)$ to create a "pseudo" continuous noisy measurements to improve recovery performance.

Algorithm 1 uDGP via projected gradient descent

Input: $\{P(y), F(y), N, \eta, 0 < \gamma < 1\}$
1: Initialize $x_0$;
2: for $t = \{0, 1, \ldots\}$ do
3: Compute the steepest gradient direction $\nabla J(x_t)$.
4: while true do
5: $x_{t+1} = \mathcal{P}(x_t - \eta \cdot \nabla J(x_t))$
6: if $J(x_{t+1}) \leq J(x_t)$ then
7: $\eta = \frac{1}{\gamma} \cdot \eta$;
8: break;
9: else
10: $\eta = \gamma \cdot \eta$
11: if $x_{t+1}$ reaches convergence then
12: $\hat{x} = x_{t+1}$;
13: break;
14: Return $\hat{x}$;

The uDGP is recasted as the following constrained nonconvex optimization problem with respect to the indicator vector $x$:

$$\min_{x} \quad J(x) = J(P(y), Q(y))$$
subject to $\|x\|_1 = N$
$$0 \leq x_m \leq 1, \forall m \in [M].$$

2.1 Matching measures via projected gradient descent

(1.3) can be solved using the projected gradient descent method given in Algorithm 1. Since this is a nonconvex problem, a proper initialization is needed to obtain a good solution. Specifically, we have the following initialization strategies:

- **Probabilistic initialization.** Without any prior information, all the 1-entries in $A_y$ have equal probabilities to correspond to the true sample pair(s) that produce(s) the distance measurement $y$. For the entry $A_{i,j}(y) = 1$, the probability that the true sample pair occupies the $i$-th cell $c_i$ and $j$-th cell $c_j$ is then: $p_y(c_i, c_j) = \frac{1}{|A_y|^2} P(y)$. The probability that the $i$-th
cell $c_j$ houses a sample can be computed by summing over both $y$ and $j$:

$$p(c_i) = \sum_{y \in R} \sum_{j=1}^{M} p_y(c_i, c_j).$$

(2.11)

We now have a probability vector $\mathbf{p}_c = \{p(c_i) | i \in [M]\}$ to tell us the likelihood of the sample locations in the domain space $\mathbb{R}^D$. Since there are $N$ sample points in total, $\mathbf{x}_0 = \mathbf{p}_c \cdot N$ naturally serves as an initialization for $\mathbf{x}$.

- **Spectral initialization.** Putting the sample pair probabilities $p_y(c_i, c_j)$ together, we have the following probability matrix $\mathbf{P}_c$:

$$\mathbf{P}_c = \sum_{y \in R} \frac{1}{\|\mathbf{A}_y\|_F^2} P(y) \cdot \mathbf{A}_y$$

(2.12)

The maximum likelihood initialization can be obtained by maximizing $\mathbf{x}^T \cdot \log \mathbf{P}_c \cdot \mathbf{x}$ subject to the constraints in (2.8) and (2.10). However, this gives rise to another difficult nonconvex optimization problem. On the other hand, suppose the samples are i.i.d in the cells following some $\mathbf{p}_c$. $p_y(c_i, c_j)$ can then be written as follows:

$$p_y(c_i, c_j) = p(c_i)p(c_j)$$

(2.13)

We can try to find such a probability vector $\hat{\mathbf{p}}_c$ so that the resulting probability matrix $\hat{\mathbf{P}}_c$ is as close to the observed $\mathbf{P}_c$ as possible. Using the Hellinger distance as the objective function, $\hat{\mathbf{p}}_c = \hat{s}^2$ can be computed as follows:

$$\hat{s} = \arg \max_s \mathbf{s}^T \sqrt{\mathbf{P}_c} \mathbf{s} \quad \text{subject to} \quad \|\mathbf{s}\|_2 = 1.$$ 

(2.14)

$\hat{s}$ is actually the leading eigenvector of $\sqrt{\mathbf{P}_c}$. The initialization vector is then: $\mathbf{x}_0 = \hat{s}^2 \cdot N$.

- **l1-min initialization.** The estimated distance frequency $\mathbf{x}^T \mathbf{A}_y \mathbf{x}$ can be rewritten as the inner product of two vectors:

$$\mathbf{x}^T \mathbf{A}_y \mathbf{x} = \text{Tr}(\mathbf{A}_y\mathbf{x}\mathbf{x}^T) = \text{Tr}(\mathbf{A}_y \mathbf{X}) = \langle \mathbf{r}_y, \mathbf{s} \rangle,$$ 

(2.15)

where $\mathbf{r}_y \in \mathbb{R}^{M^2}$ is obtained by concatenating every row of $\mathbf{A}_y$, and $\mathbf{s} \in \{0, 1\}^{M^2}$ is a sparse vector obtained by concatenating every column of $\mathbf{X} = \mathbf{x}\mathbf{x}^T$. (2.5) can be relaxed to the following $l_1$-minimization problem:

$$\hat{s} = \arg \min_s \|\mathbf{f} - \mathbf{R}\mathbf{s}\|_2$$

subject to $\|\mathbf{s}\|_1 = N^2$

(2.16)

where $\mathbf{f} \in \mathbb{R}^{\left|\mathbb{R}\right|}$ is a vector containing all the observed distance frequency measures $F(y)$; $\mathbf{R} \in \{0, 1\}^{M^2 \times M^2}$ is a sparse matrix whose every row is a vectorized $\mathbf{A}_y$, i.e. $\mathbf{r}_y$; $\lambda > 0$ is the regularization parameter. (2.16) is a convex problem, and can be solved efficiently. From $\hat{s}$ we can construct a matrix $\hat{\mathbf{X}}$. The initializer $\mathbf{x}_0$ can be obtained by finding the rank-one approximation of $\hat{\mathbf{X}}$:

$$\mathbf{x}_0 = \arg \min_{\mathbf{x}} \|\hat{\mathbf{X}} - \mathbf{x}\mathbf{x}^T\|_2^2.$$ 

(2.17)
Figure 5: In projected gradient descent, the gradient descent update $x = x_t - \eta \cdot \nabla J(x_t)$ is projected back to the convex set $S$: $s = \arg \min_{s \in S} \|s - x\|_2^2$.

2.2 Efficient projection onto the $l_1$-ball with box constraints

Algorithm 1 uses projected gradient descent to search for a local optimal solution. As is shown in Figure 5, every time after the gradient descent update $x = x_t - \eta \cdot \nabla J(x_t)$, $x$ needs to be projected back onto the $l_1$-ball with box constraints (2.8) and (2.10): $s = \mathcal{P}(x_t - \eta \cdot \nabla J(x_t))$. The projection is done through the following optimization problem:

$$
\min_s \frac{1}{2} \|s - x\|_2^2 \\
\text{s.t. } \sum_{m=1}^M s_m = N \\
\quad \quad \quad 0 \leq s_m \leq 1
$$

(2.18)

[60] proposed an efficient approach for Euclidean projections where $s_m$ is lower-bounded, it solves an unconstrained optimization problem with Lagrange multipliers. [5,30] later extended [60]'s approach to handle projections with box constraints, i.e. $s_m$ is both lower-bounded and upper-bounded. However, their approach is based on the sequential search of some optimal threshold $\omega$, and thus couldn’t be parallelized for large-scale problems.

Based on the work of [60], here we take advantage of the properties of the optimal solution and develop an efficient parallelizable approach to obtain projections with box constraints. Specifically, the Lagrangian of (2.18) is:

$$
\mathcal{L} = \frac{1}{2} \|s - x\|_2^2 + \omega \left( \sum_{m=1}^M s_m - N \right) - \zeta^T \cdot s + \xi^T \cdot (s - 1),
$$

(2.19)

where $\omega \in \mathbb{R}$, $\zeta \in \mathbb{R}_+^M$, $\xi \in \mathbb{R}_+^M$ are the Lagrange multipliers. Taking the derivative w.r.t. $s$, and setting it to 0, we have $\frac{\partial \mathcal{L}}{\partial s_m} = s_m - x_m + \omega - \zeta_m + \xi_m = 0$. The complementary slackness KKT
condition tells us that when $0 \leq s_m \leq 1$, $\zeta_m = \xi_m = 0$. We then have:

$$s_m = x_m - \omega \quad \text{if } 0 \leq s_m \leq 1. \quad (2.20)$$

[60] has the following lemma on the lower bound $0 \leq s_m$ about the optimal solution $s$:

**Lemma 2.2** [60] Let $s$ be the optimal solution to the minimization problem in (2.18). Let $i$ and $j$ be two indices such that $s_i > s_j$. If $s_i = 0$ then $s_j$ must be 0 as well.

Similarly we can write the following lemma on the upper bound $s_m \leq 1$:

**Lemma 2.3** Let $s$ be the optimal solution to the minimization problem in (2.18). Let $i$ and $j$ be two indices such that $s_i > s_j$. If $s_j = 1$ then $s_i$ must be 1 as well.

**Proof** Since $s_i \leq 1$, we first assume that $s_i < 1$. We construct a vector $\tilde{s} \in \mathbb{R}^M$ out of $s$ by switching the positions of $s_i$ and $s_j$ in $s$, i.e. $\tilde{s}_i = s_j$ and $\tilde{s}_j = s_i$. $\tilde{s}$ also satisfies the constraints in (2.18). We then have:

$$\|s - x\|^2 - \|\tilde{s} - x\|^2 = (s_i - x_i)^2 + (1 - x_j)^2 - (1 - x_i)^2 - (s_i - x_j)^2 = 2(1 - s_i)(s_i - s_j) \geq 0. \quad (2.21)$$

This is in contradiction with the fact that $s$ is the minimizer of (2.18). Hence $s_i$ must be 1. 

Since reordering of the entries of $x$ doesn’t change the optimization process of (2.18), and adding some constant $\tau$ to $x$ doesn’t change the solution of (2.18), without loss of generality we are going to assume that the entries of $x$ are all positive in a non-increasing order: $x_1 \geq x_2 \geq \cdots \geq x_M \geq N$.

Lemma 2.2 and 2.3 imply that for the minimizing $s$:

- The entries of $s$ are in a non-decreasing order.
- The first $\rho$ entries of $s$ are between 0 and 1, i.e. $0 < s_m \leq 1$; the rest entries are all 0s.

Additionally, there must exists some $r$: $1 \leq r \leq \rho$ such that:

$$1 < x_{r-1} - \omega \quad (2.22)$$

$$0 < x_r - \omega \leq 1 \quad (2.23)$$

We can write the sum of $s$ as follows:

$$\sum_{m=1}^{M} s_m = \sum_{m=1}^{\rho} s_m = \sum_{m=1}^{r-1} 1 + \sum_{m=r}^{\rho} (x_m - \omega) = N. \quad (2.24)$$

We then have:

$$\omega = \frac{1}{\rho - r + 1} \left( \sum_{m=r}^{\rho} x_m - (N - r + 1) \right). \quad (2.25)$$

The minimizing $s$ is then:

$$s = \begin{cases} 
1, & m \leq r - 1 \\
 x_m - \omega, & r - 1 < m \leq \rho \\
0, & \rho < m \leq M.
\end{cases} \quad (2.26)$$
Algorithm 2 Efficient projection onto the $l_1$-ball with box constrains

1: Shift $x = x + \tau$ s.t. $x_m \geq N, \forall m.$
2: $v = \text{sort}(x)$.
3: for $r = \{1, 2, \cdots, M\}$ do
4: \hspace{1em} Construct $\hat{v}$ out of $v$ by removing the first $r - 1$ entries.
5: \hspace{1em} $\rho_v = \max \left\{ l \in [N - r + 1] : \hat{v}_l - \frac{1}{l} \left( \sum_{m=1}^{l} \hat{v}_m - (N - r + 1) \right) \right\}$.
6: \hspace{1em} $\omega_v = \frac{1}{\rho_v} \left( \sum_{m=1}^{\rho_v} \hat{v}_m - (N - r + 1) \right)$.
7: \hspace{1em} Check the candidate $\omega_v$: $\hat{u}_r = \max(x_r - \omega_v, 0)$.
8: \hspace{1em} if $0 < \hat{u}_r \leq 1$ then
9: \hspace{2em} $\omega = \omega_v$;
10: \hspace{2em} break;
11: \hspace{1em} else
12: \hspace{2em} $\hat{u}_{r-1} = \max(x_{r-1} - \omega_v, 0)$.
13: \hspace{2em} if $\hat{u}_{r-1} > 1$ then
14: \hspace{3em} $\omega = \omega_v$;
15: \hspace{3em} break;
16: \hspace{2em} else
17: \hspace{3em} continue;
18: \hspace{1em} end if
19: $s = \max(x - \omega, 0)$ and $s = \min(s, 1)$.
20: Return $s$.

If we can find the value of $r$ and $\rho$, the Lagrange multiplier $\omega$ can be computed using (2.25) and the problem is solved. Suppose that the value of $r$ is known, the search for $\omega$ and $\rho$ can then be efficiently completed using the approach by [60].

However, since $r$ is unknown in practice, we can check each $r \in [M]$ one by one to see if the computed $(\rho, \omega|r)$ satisfy the above two constraints (2.22) and (2.23). Specifically, we have the following lemma:

Lemma 2.4 There is one and only one $r \in [M]$ that produces the $(\rho, \omega|r)$ that satisfy (2.22) and (2.23).

Proof Please refer to Appendix A.2. \hfill \blacksquare

Based on lemma 2.4, we next propose the following Algorithm 2 to find the values of $r$ and $\rho$.

2.3 Local convergence of the projected gradient descent method

Let $S$ denote the convex set defined by (2.8) and (2.10); $\hat{S}$ denote the convex set defined by (2.8), (2.10), and $x^T A_y x \neq 0, \forall y \in \mathcal{R}$. Hence $\hat{S} \subset S$. Taking the first-order derivative of the objective functions in (2.5)-(2.7) with respect to $x$, since $A_y$ is a symmetric matrix, we have:

• Squared Euclidean distance:

\[ \nabla J(x) = \sum_{y \in \mathcal{R}} (x^T A_y x - F(y)) \cdot A_y x, \quad \forall x \in S. \tag{2.27} \]
• Hellinger distance:

$$\nabla J(x) = \sum_{y \in \mathcal{R}} - \sqrt{\frac{F(y)}{x^T A_y x}} \cdot A_y x, \quad \forall x \in \hat{S}. \quad (2.28)$$

• Cross entropy:

$$\nabla J(x) = \sum_{y \in \mathcal{R}} - \frac{F(y)}{x^T A_y x} \cdot A_y x, \quad \forall x \in \hat{S}. \quad (2.29)$$

As is shown in Appendix A.3, it is easy to verify that the second-order derivative of squared Euclidean distance (2.5) is bounded in the convex set $\mathcal{S}$. Hence (2.27) is Lipschitz continuous in $\mathcal{S}$.

For the hellinger distance and cross entropy, if $\nabla^2 J(x)$ is bounded, there is a small neighbourhood $\rho(x)$ around $x$ such that $\forall s \in \rho(x)$, $\nabla^2 J(s)$ is also bounded. Hence (2.28) and (2.29) are locally Lipschitz continuous.

The backtracking strategy is used in Algorithm 1 to determine the step size $\eta$ of the projected gradient descent method. We can always find a valid step size $\eta > 0$ using backtracking:

**Lemma 2.5** If $\nabla^2 J(x_t)$ is bounded, there exists some $\eta > 0$ such that $J(s) \leq J(x_t)$, where $s = \mathcal{P}(x_t - \eta \cdot \nabla J(x_t))$.

**Proof** Please refer to Appendix A.4.

The objective function $J(x_t)$ is thus monotonically non-increasing with respect to the solution $x_t$. Combined with the fact that $J(x)$ is lower-bounded by the global minimum value, we can see that Algorithm 1 will eventually reach local convergence.

### 3 Convergence to the global optimum

#### 3.1 Squared Euclidean distance

Taking the squared Euclidean distance in (2.5) as the objective function, we have the following optimization problem:

$$\hat{x} = \arg \min_x \sum_{y \in \mathcal{R}} \frac{1}{4} \left( x^T A_y x - x^* T A_y x^* \right)^2. \quad (2.5 \text{ revisited})$$

##### 3.1.1 Noiseless recovery

In the noiseless case, as is illustrated in Fig. 6(a), we can show that there is a small neighbourhood $\rho(x^*)$ around every global optimizer $x^* \in \{0, 1\}^M$, such that if $\|x_t - x^*\|_2 \leq \tau_{\text{max}}$, the projected gradient descent update $x_{t+1} = \mathcal{P}(x_t - \eta \cdot \nabla J(x_t))$ is guaranteed to converge to $x^*$, as is shown in the Theorem 3.2. First, we have the following lemma 3.1:
Lemma 3.1  Let $x^*$ be a global optimizer, and $E = \sum_{y \in \mathbb{R}} A_y x^* x^*^T A_y^T$, then $\forall r \in S, r \neq x^*$ we have the following convex problem:

$$
\lambda(E) = \min_{r \in S, r \neq x^*} \frac{1}{\|x - x^*\|_2^2}(r - x^*)^T E (r - x^*)
$$

$$
= \min_{z \in Z} z^T Ez
$$

$$
> 0
$$

where $Z$ is a convex set defined by the following constraints:

$$
\|z\|_1 = f^T z = 1
$$

$$
\sum_{i=1}^M z_i = 0
$$

$$
z_i \in [0, 0.5] \text{ if } x_i^* = 0
$$

$$
z_i \in [-0.5, 0] \text{ if } x_i^* = 1
$$

$f \in \{-1, 1\}^M$ depends on $x^*$ and is as follows:

$$
f_i = \begin{cases} 
1 & \text{if } x_i^* = 0 \\
-1 & \text{if } x_i^* = 1
\end{cases}
$$

Proof  Please refer to Appendix A.5. 

Using Lemma 3.1, we can prove our main theorem for the noiseless case:

**Theorem 3.2**  Given an initialization $x_t \in \mathbb{R}^M$ within a small neighbourhood $\rho(x^*) \subset S$ of a global optimizer $x^*$ s.t.

$$
\|h\|_2 = \|x_t - x^*\|_2 = \tau < \tau_{\text{max}} = \left(2 - \frac{1}{q}\right) \cdot \sqrt{\frac{1}{\sigma_{\text{max}}^2}(E)}
$$

where $\frac{1}{q} < q < 1$; $\sigma_{\text{max}}$ is the largest maximum singular value of all $A_y$: $\sigma_{\text{max}} \geq \sigma_{\text{max}}(A_y)$; $E = \sum_{y \in \mathbb{R}} A_y x^* x^*^T A_y^T$.

The projected gradient descent minimization of $J(x) = \sum_{y} \frac{1}{2} \left( x^T A_y x - x^* T A_y x^* \right)^2$ is guaranteed to converge to $x^*$ when the step size $\eta$ satisfies: $0 < \frac{1-q}{\|x_t\|_2^2} - \nu < \eta < \frac{1-q}{\|x_t\|_2^2} + \nu$:

$$
\|x_{t+1} - x^*\|_2^2 < \mu \|x_t - x^*\|_2^2
$$

where $\max \left\{ 0, 1 - \frac{(1-q)^2}{\|x_t\|_2^2} \frac{4J(x_t)}{\sigma_{\text{max}}^2} \right\} < \mu < 1$, and $\nu = \sqrt{\frac{(1-q)^2}{\|x_t\|_2^2} \frac{4J(x_t)}{\sigma_{\text{max}}^2} - \frac{(1-\mu)^2}{\|x_t\|_2^2} J(x_t)}$.

Proof  Please refer to Appendix A.6.
In the noisy case, $\lambda(E) \geq 0$ in (3.1). As is illustrated in Fig. 6(b), there is not necessarily a small neighbourhood around every global optimizer $x^* \in [0,1]^M$ that guarantees the convergence to $x^*$. However, there do exist regions next to $x^*$ that support the convergences to $x^*$. As is given in Appendix A.6, such regions $V$ next to $x^*$ can be decided as follows:

$$V = \left\{ x \left| \|x - x^*\|_2 < \left( 2 - \frac{1}{q} \right) \right. \cdot \frac{z^T E z}{\|x - x^*\|_1} \right. \left. \text{ and } x \in S \right\} \tag{3.9}$$

- If the intersection of $Z$ and the null space of $E$ is empty: $Z \cap \text{Null}(E) = \emptyset$, $\lambda(E) > 0$ and the convergence neighbourhood around $x^*$ exists, as is illustrated in Fig. 6(a).
- If $Z \cap \text{Null}(E) \neq \emptyset$, $\lambda(E) \geq 0$. There exist regions defined by (3.9) that support the convergence to $x^*$, as is illustrated in Fig. 6(b).

### 3.2 Hellinger distance

Choosing the Hellinger distance as the objective function, we have the following optimization problem:

$$\hat{x} = \arg \min_x \sum_{y \in \mathcal{R}'} -\sqrt{F(y) \cdot x^T A_y x}. \tag{2.6 revisited}$$

where $\mathcal{R}' = \{y \mid F(y) > 0\}$ and $\mathcal{R}' \subset \mathcal{R}$. 

---

Figure 6: Convergence behavior to a global optimizer $x^*$ of the projected gradient descent method: (a) the neighbourhood $\{x \mid \|x - x^*\|_2 < \tau_{\text{max}}\}$ guarantees the convergence if $Z \cap \text{Null}(E) = \emptyset$; (b) the region defined by (3.9) next to $x^*$ if $Z \cap \text{Null}(E) \neq \emptyset$. 

---

| convergence neighbourhood around $x^*$ | convergence region next to $x^*$ |
|----------------------------------------|----------------------------------|
| ![Diagram](image1.png)                | ![Diagram](image2.png)            |

---

(3.9)
3.2.1 Noiseless recovery

There exists a small neighbourhood \( \rho(x^*) \) around every global optimizer \( x^* \) that guarantees the convergence to it using the projected gradient descent method, as is given by the following Theorem 3.3.

**Theorem 3.3** Given an initialization \( x_0 \in \mathbb{R}^M \) within a small neighbourhood \( \rho(x^*) \subset \tilde{S} \) of a global optimizer \( x^* \) s.t.

\[
\| h \|_2 = \| x_t - x^* \| < \tau_{\text{max}} = \min \left\{ \phi, \left( 2 - \frac{1}{q} \right) \cdot \sqrt{\frac{1}{\sigma_{\text{max}} \Lambda_{\text{min}}} \lambda(E)} \right\},
\]

where \( \frac{1}{2} < q < 1; \phi < x^*_{\min}, \lambda_{\min} \) is the smallest nonzero entry of \( x^* \); \( \sigma_{\text{max}} \) is the largest maximum singular value of all \( A_y \): \( \sigma_{\text{max}} \geq \sigma_{\text{max}}(A_y) \); \( E = \sum_{y \in \mathcal{R'}} A_y x^* x^T A_y^T \); \( \kappa_{\min} \) and \( \kappa_{\max} \) are given in (A.64),(A.65).

The projected gradient descent minimization of \( J(x) = \sum_{y \in \mathcal{R'}} -\sqrt{F(y)} \cdot x^T A_y x \) is guaranteed to converge to \( x^* \) when the step size \( \eta \) satisfies: \( 0 < \frac{(1-q)\Gamma_1}{\|E\|_{\sigma_{\text{max}},x_t}^2 \Gamma_1} - \nu < \eta < \frac{(1-q)\Gamma_2}{\|E\|_{\sigma_{\text{max}},x_t}^2 \Gamma_1} + \nu \),

where \( \Gamma_1 = \sum_y \frac{F(y)}{x_t^T A_y x_t} \) and \( \Gamma_2 = \sum_y \frac{((x_t - x^*)^T A_y (x_t + x^*))^2}{\sqrt{x_t^T A_y x_t} \sqrt{x_t^T A_y x_t + \sqrt{F(y)}}} \).

\[
\| x_{t+1} - x^* \|_2^2 < \mu \| x_t - x^* \|_2^2,
\]

where \( \max \left\{ 0, 1 - \frac{(1-q)^2 \Gamma_2^2}{\|E\|_{\sigma_{\text{max}},x_t}^2 \Gamma_1 \tau_{\text{max}}^2} \right\} < \mu < 1 \), and \( \nu = \sqrt{\frac{(1-q)^2 \Gamma_2^2}{\|E\|_{\sigma_{\text{max}},x_t}^2 \Gamma_1 \tau_{\text{max}}^2}} - \frac{(1-\mu)\Gamma_2^2}{\|E\|_{\sigma_{\text{max}},x_t}^2 \Gamma_1 \tau_{\text{max}}^2} \).

**Proof** Please refer to Appendix A.7.

3.2.2 Noisy recovery

- If the intersection of \( \mathcal{Z} \) and the null space of \( E \) is empty: \( \mathcal{Z} \cap \text{Null}(E) = \emptyset \), \( \lambda(E) > 0 \) and the convergence neighbourhood around \( x^* \) exists, as is illustrated in Fig. 6(a). Furthermore, we can see from (3.12) that the size of the convergence neighbourhood is determined by the smallest nonzero entry of \( x^* \): \( x^*_{\min} \). In the noisy case \( x^*_{\min} \) is usually much less than 1.

- If \( \mathcal{Z} \cap \text{Null}(E) \neq \emptyset \), \( \lambda(E) \geq 0 \). There exist regions defined by (3.9) that support the convergence to \( x^* \), as is illustrated in Fig. 6(b).

3.3 Cross entropy

The optimization problem using the cross entropy as the objective function is as follows:

\[
\hat{x} = \arg \min_x \sum_{y \in \mathcal{R'}} -\frac{1}{2} F(y) \log x^T A_y x,
\]

where \( \mathcal{R}' = \{ y \mid F(y) > 0 \} \) and \( \mathcal{R}' \subset \mathcal{R} \).
3.3.1 Noiseless recovery

The small neighbourhood around every global optimizer $x^*$ that guarantees the convergence to it using the projected gradient update is given by the following Theorem 3.4.

Theorem 3.4 Given an initialization $x_t \in \mathbb{R}^M$ within a small neighbourhood $\rho(x^*) \subset \hat{S}$ of a global optimizer $x^*$ s.t.

$$\|h\|_2 = \|x_t - x^*\| < \tau_{\max} = \min\left\{ \phi, \left(2 - \frac{1}{q}\right) \cdot \sqrt{\frac{1}{\sigma_{\max} \kappa_{\min}}} \lambda(E) \right\},$$

where $\frac{1}{2} < q < 1$; $\phi < x^*_{\min}$; $x^*_{\min}$ is the smallest nonzero entry of $x^*$; $\sigma_{\max}$ is the largest maximum singular value of all $A_y$; $\sigma_{\max} \geq \sigma_{\max}(A_y)$; $E = \sum_{y \in \mathcal{R}'} A_y x^* x^T A^T_y$; $\kappa_{\min}$ and $\kappa_{\max}$ are given in (A.82), (A.83).

The projected gradient descent minimization of $J(x) = \sum_{y \in \mathcal{R}'} -\frac{1}{2} F(y) \log x^T A_y x$ is guaranteed to converge to $x^*$ when the step size $\eta$ satisfies: $0 < \frac{(1-q) \Gamma_2}{|R|\|\sigma_{\max}\|_{\mathbf{x}_t}^2 \Gamma_1} - \nu < \frac{(1-q) \Gamma_2}{|R|\|\sigma_{\max}\|_{\mathbf{x}_t}^2 \Gamma_1} + \nu$, where $\Gamma_1 = \sum_y \frac{F(y)}{x^T A_y x_t}$ and $\Gamma_2 = \sum_y \left(\frac{(x_t - x^*)^T A_y (x_t + x^*)}{x^T A_y x_t}\right)^2$.

$$\|x_{t+1} - x^*\|_2^2 < \mu \|x_t - x^*\|_2^2,$$

where $\max\left\{0, 1 - \frac{(1-q)^2 \Gamma_2^2}{|R|\|\sigma_{\max}\|_{\mathbf{x}_t}^2 \Gamma_1 \tau_{\max}}\right\} < \mu < 1$, and $\nu = \sqrt{\frac{(1-q)^2 \Gamma_2^2}{|R|\|\sigma_{\max}\|_{\mathbf{x}_t}^2 \Gamma_1}} - \frac{(1-q) \Gamma_2^2}{|R|\|\sigma_{\max}\|_{\mathbf{x}_t}^2 \Gamma_1}$.

Proof Please refer to Appendix A.8.

3.3.2 Noisy recovery

The analysis of the noisy recovery case is the same as discussed for the Hellinger distance objective function in section 3.2.2.

4 Numerical experiments

The proposed distance distribution matching for the uDGP in (1.3) is nonconvex, and several initialization approaches are introduced in section 2.1. Here we choose the spectral initialization approach for the simulation experiments. Additionally, the choice of objective functions also affect the recovery performance. For instance, the earth mover’s distance fails to deliver a decent recovery when the number of points is larger than 10. Here we shall conduct 1D noisy recovery experiments to evaluate the performance of the proposed distance distribution matching approach with different choices of objective function (2.5)-(2.7).

In the noisy case, the gradients of the Hellinger distance (2.28) and the cross entropy (2.29) could become very large for some estimated frequency $x^T A_y x_t$ during the optimization process, which causes the projected gradient descent update to be trapped in some local optimum. To avoid such situations, we only keep the gradients at certain distance values $y \in \hat{R}_t$ if $x^T A_y x_t$ is larger than some pre-specified threshold $\kappa$, where $\hat{R}_t$ is as follows:

$$\hat{R}_t = \{ y \mid x^T A_y x_t \geq \kappa > 0 \}.$$
Note that $\hat{R}_t$ will be updated at every iteration.

We randomly sample $N = 100$ points from the interval $[0, 1]$ with the minimum pairwise distance set to $1e^{-4}$ and the maximum pairwise distance set to 1. The distance measurement $y$ is further corrupted by an additive white Gaussian noise: $y_{ui} = |u_i - u_j| + w$, where $i < j$ and $w \sim \mathcal{N}(0, \sigma^2)$. Generally speaking, a successful recovery require that the noise is smaller than the minimum pairwise distance. We thus control the amount of noise added to the measurements by varying the standard deviation $\sigma$: $\sigma \in \{1e^{-5}, 3e^{-5}, 5e^{-5}, 9e^{-5}\}$.

For the Hellinger distance and cross entropy objective functions, the $\kappa$ in (4.1) is chosen to be $1e^{-3}$ for best performance. Since we do not know the true standard deviation of the added noise, we try different test standard deviations $\hat{\sigma}$ during the recovery: $\hat{\sigma} \in \{1e^{-5}, 3e^{-5}, 5e^{-5}, 9e^{-5}\}$.

The recovered point set $\hat{\mathcal{U}}$ can be matched to the true point set $\mathcal{U}$ efficiently using the Hungarian algorithm [39], and the average assignment cost per point is shown in Fig. 7. If the assignment cost of a point is less than the smallest pairwise distance $1e^{-4}$, the point is considered to be recovered correctly, and the number of correctly recovered points is recorded. For every choice of $\{N, \sigma, \hat{\sigma}\}$, 100 random runs is performed and the average number of correctly recovered points is reported as the final results, as is shown in Fig. 8. We can see that the proposed distance distribution matching approach is robust to the white Gaussian noise with standard deviation $\sigma$ up to $5e^{-5}$, whereas conventional greedy build-up approaches become inoperable. In fact, $\sigma = 5e^{-5}$ is roughly the threshold where the noise $w$ starts to be greater than the smallest pairwise distance $1e^{-4}$.

5 Conclusion

In this paper we recast the unassigned distance geometry problem into a constrained nonconvex optimization problem given in (1.3). The locations of the sample points can be obtained by searching for an indicator vector $x$ so that the estimated distance distribution produced by $x$ matches the observed distance distribution. Loss functions such as squared Euclidean distance, Hellinger distance, cross entropy, can be used to quantify the difference between the estimated and observed
distance distribution. In an effort to efficiently solve (1.3), several initialization schemes are introduced to properly initialize the proposed projected gradient descent method. Furthermore, we also outline the necessary conditions for projected gradient descent update to converge to the global optimum $x^*$ in the noiseless and noisy cases. Simulated numerical experiments show that the proposed distance distribution matching approach is robust to the noise. This opens the door to the possibilities of applying it on real noisy measurements and helping practical applications such as the nanostructure problem, NMR spectroscopy, etc.

Figure 8: The average number of successfully recovered points from 100 random runs using different objective functions. In each random run, 100 points are randomly sampled from the interval $[0,1]$ with the minimum pairwise distance set to $1e^{-4}$ and the maximum pairwise distance set to 1.
A  Proofs

A.1 Lemma 2.1

Proof  Since (2.5)-(2.7) are minimized if and only if \( Q(y) = P(y) \). A global optimizer \( x \) must satisfy \( Q(y = 0) = P(y = 0) \):

\[
x^TA_0x = x^Tx = \sum_{m=1}^{M} x_m^2 = N.
\]

(A.1)

From (2.8), we have:

\[
\sum_{m=1}^{M} x_m^2 = \sum_{m=1}^{M} x_m.
\]

(A.2)

From (2.10), we have

\[
x_m^2 \leq x_m, \quad \forall m = [M].
\]

(A.3)

Suppose the entry \( x_m \in (0, 1) \), then \( x_m^2 \) is strictly less than \( x_m \). Combining \( x_m^2 < x_m \) with (A.3), we have \( \sum_{m=1}^{M} x_m^2 < \sum_{m=1}^{M} x_m \). This contradicts (A.2), hence \( x_m \notin (0, 1) \). From (2.10) we know that \( x_m \in [0, 1] \). Hence \( x_m \in \{0, 1\} \). □

A.2 Lemma 2.4

Proof  We will proceed in the following two steps:

1. Since the set \( S_1 \) is a non-empty set, the projection onto it exists, i.e. there is one \( r \in [M] \) that produces the \( (\rho, \theta) \) that satisfy (2.22) and (2.23). In fact, since \( S_1 \) is a closed convex, the projection is also unique.

2. With out loss of generality, suppose that there are two different \( r_1 < r_2 \in [M] \) that produce the two pairs \( (\rho_1, \theta_1 | r_1) \) and \( (\rho_2, \theta_2 | r_2) \) that satisfy the constraints. We have:

\[
r_1 < r_2 \Rightarrow r_1 < r_2 - 1 \Rightarrow x_{r_2-1} - \theta_1 < x_{r_1} - \theta_1 \Rightarrow x_{r_2-1} - 1 < \theta_1
\]

\[
1 < x_{r_2-1} - \theta_2 \Rightarrow \theta_2 < x_{r_2-1} - 1
\]

Hence \( \theta_2 < \theta_1 \).

\[
x_{\rho_1} - \theta_1 > 0 \Rightarrow x_{\rho_1} > \theta_1
\]

\[
x_{\rho_2+1} - \theta_2 \leq 0 \Rightarrow x_{\rho_2+1} \leq \theta_2
\]

\[
x_{\rho_2+1} \leq \theta_2 < \theta_1 < x_{\rho_1} \Rightarrow x_{\rho_2+1} < x_{\rho_1}
\]

Hence \( \rho_2 + 1 > \rho_1 \Rightarrow \rho_2 \geq \rho_1 \).
(a) If \( r_2 \leq \rho_1 \), we have:

\[
\begin{align*}
\rho_1 - 1 + \sum_{m=r_1}^{\rho_1} (x_m - \theta_1) &= \rho_1 - 1 + \sum_{m=r_1}^{\rho_2} (x_m - \theta_1) + \sum_{m=r_2}^{\rho_2} (x_m - \theta_1) \\
&\leq \rho_1 - 1 + \sum_{m=r_1}^{\rho_1} (x_m - \theta_1) + \sum_{m=r_2}^{\rho_2} (x_m - \theta_1) \\
&< \rho_1 - 1 + \sum_{m=r_2}^{\rho_1} (x_m - \theta_2) \\
&\leq \rho_1 - 1 + \sum_{m=r_2}^{\rho_2} (x_m - \theta_2)
\end{align*}
\]

\[(A.4)\]

(b) If \( r_2 > \rho_1 \), we have:

\[
\begin{align*}
\rho_1 - 1 + \sum_{m=r_1}^{\rho_1} (x_m - \theta_1) &\leq \rho_1 - 1 + \sum_{m=r_1}^{\rho_1} 1 \\
&= \rho_1 \\
&\leq \rho_1 - 1 \\
&< \rho_1 - 1 + \sum_{m=r_2}^{\rho_2} (x_m - \theta_2)
\end{align*}
\]

\[(A.5)\]

Both (A.4) and (A.5) are in contradiction with the fact that the solutions \( s_1, s_2 \) produced by \( r_1, r_2 \) have equal sums. Hence \( r_1 = r_2 \), there is only one \( r \) that produces the \((\rho, \theta)\) that satisfy the constraints.

\[
\blacklozenge
\]

### A.3 Second order gradients of the objective functions

- Squared Euclidean distance:

\[
\nabla^2 J(x) = \sum_{y \in \mathcal{R}} \left[ 2A_y xx^T A_y^T + (x^T A_y x - F(y)) \cdot A_y \right]
\]

\[(A.6)\]

- Hellinger distance:

\[
\nabla^2 J(x) = \sum_{y \in \mathcal{R}} \sqrt{\frac{F(y)}{x^T A_y x}} \left[ \frac{A_y xx^T A_y^T}{x^T A_y x} - A_y \right]
\]

\[(A.7)\]

- Cross entropy:

\[
\nabla^2 J(x) = \sum_{y \in \mathcal{R}} \frac{F(y)}{x^T A_y x} \left[ \frac{2A_y xx^T A_y^T}{x^T A_y x} - A_y \right]
\]

\[(A.8)\]

20
A.4 Lemma 2.5

Proof Since \( x_t \in S \), we can represent any point \( x \in \mathbb{R}^M \) in the convex set \( S \) by
\[
x = x_t + r,
\]
where \( \sum_{m=1}^{M} r_m = 0 \) and \( -x_{tm} < r_m < 1 - x_{tm}, \forall m \in [M] \).

1. To simplify notation, we use \( J' \) to denote the gradient \( \nabla J(x_t) \). The projection \( \mathcal{P}_S(\cdot) = \) is then:
\[
\begin{align*}
\arg\min_{x \in S} \| x - (x_t - \eta \cdot J') \|_2^2 \iff \arg\min_r \| r + \eta \cdot J' \|_2^2
\end{align*}
\]
J' can be decomposed into the sum of two vectors \( J'_1 \) and \( J'_2 \): \( J' = J'_1 + J'_2 \). Let \( 1 \in \mathbb{R}^M \) denote a vector of all 1s, we have:
\[
\begin{align*}
J'_1 &= \frac{1^T J'}{M} \cdot 1 \\
J'_2 &= J' - J'_1
\end{align*}
\]
\( J'_2 \) is essentially the gradient of \( J(x) \) in the convex set \( S \). Since \( r^T J'_1 = 0 \), we get:
\[
\begin{align*}
\hat{r} &= \arg\min_r \| r + \eta \cdot (J'_1 + J'_2) \|_2^2 \\
&= \arg\min_r r^T r + 2\eta \cdot r^T J'_2
\end{align*}
\]
Since we can always make \( r^T r + 2\eta \cdot r^T J'_2 = 0 \) by setting \( r = 0 \), we then have the following inequality for the minimizing \( \hat{r} \):
\[
\hat{r}^T \hat{r} + 2\eta \cdot \hat{r}^T J'_2 \leq 0
\]
Since \( \hat{r}^T \hat{r} \geq 0 \) and \( \eta > 0 \), we must have:
\[
\hat{r}^T J'_2 \leq 0
\]
2. Since \( \nabla^2 J(x_t) \) is bounded, \( \nabla J(x_t) \) is at least locally Lipschitz continuous in the small neighbourhood \( \rho(x_t) \) around \( x_t \). In other words, \( \forall s \in \rho(x_t), J(s) \) has Lipschitz continuous gradient. We can then obtain a quadratic upper bound on \( J(s) \). Suppose the Lipschitz constant is \( L \), we have:
\[
J(s) = J(x_t + \hat{r}) \\
\leq J(x_t) + \hat{r}^T J'_2 + \frac{L}{2} \hat{r}^T \hat{r}
\]
If the step size \( \eta \) is chosen such that \( 0 < \eta \leq \frac{1}{L} \), using (A.15), we have:
\[
\frac{2}{L} \cdot \hat{r}^T J'_2 \leq 2\eta \cdot \hat{r}^T J'_2
\]
Plug (A.17) into (A.14), we can get:
\[
\hat{r}^T J'_2 + \frac{L}{2} \hat{r}^T \hat{r} \leq 0
\]
Hence \( J(s) \leq J(x_t) \).

\[\blacksquare\]
A.5 Lemma 3.1

Proof Since $E = \sum_{y \in \mathcal{Y}} A_y x^* x^T A_y^T$, we can see that $s^T E s = \sum_{y} (s^T E x^*)^2 \geq 0$, $\forall s \in \mathbb{R}^M$. Hence $E$ is positive-semidefinite, and $s^T E s$ is a convex function of $s$ in $\mathbb{R}^M$. We define the following set $Z_1$:

$$Z_1 = \left\{ z \mid z = \frac{1}{\|r - x^*\|_1} (r - x^*), \forall r \in \mathcal{S}, r \neq x^* \right\} \quad \text{(A.19)}$$

where $\mathcal{S}$ is the convex set defined by (2.8) and (2.10). We next prove that $Z_1$ is a convex set. First, we can see that $\forall z^{(1)} \neq z^{(2)}$, we have $\text{sign}(z_i^{(1)}) = \text{sign}(z_i^{(2)})$, $\forall i \in [M]$. Let $\rho \in [0, 1]$, we have:

$$\left\| (1 - \rho) z^{(1)} + \rho z^{(2)} \right\|_1 = \sum_{i=1}^{M} \left| (1 - \rho) z_i^{(1)} + \rho z_i^{(2)} \right|$$

$$= \sum_{i=1}^{M} \left| (1 - \rho) z_i^{(1)} \right| + \left| \rho z_i^{(2)} \right|$$

$$= (1 - \rho) \left\| z^{(1)} \right\|_1 + \rho \left\| z^{(2)} \right\|_1 = 1 \quad \text{(A.20)}$$

Let $\nu_1 = \frac{1 - \rho}{\|r^{(1)} - x^*\|_1}$, $\nu_2 = \frac{\rho}{\|r^{(2)} - x^*\|_1}$. We also have:

$$(1 - \rho) z^{(1)} + \rho z^{(2)} = \frac{1 - \rho}{\|r^{(1)} - x^*\|_1} (r^{(1)} - x^*) + \frac{\rho}{\|r^{(2)} - x^*\|_1} (r^{(2)} - x^*)$$

$$= (\nu_1 + \nu_2) \left( \frac{\nu_1}{\nu_1 + \nu_2} r^{(1)} + \frac{\nu_2}{\nu_1 + \nu_2} r^{(2)} - x^* \right)$$

$$= (\nu_1 + \nu_2) (r^{(3)} - x^*). \quad \text{(A.21)}$$

From (A.20), we can see that $\frac{1}{\nu_1 + \nu_2} = \|r^{(3)} - x^*\|$. Since $r^{(3)} \in \mathcal{S}$, we have $(1 - \rho) z^{(1)} + \rho z^{(2)} \in Z_1$. Hence $Z_1 \subseteq \mathbb{R}^D$ is a convex set, and $z^T E z$ is a convex function of $z \in Z_1$.

Furthermore, if $z^T E z = \sum_{y} (z^T A_y x^*)^2 = 0$, we have $(r - x^*)^T A_y x^* = 0$, $\forall y \in \mathcal{R}$. When $y = 0$, $A_0 = I$ is the identity matrix, we then have $r^T x^* = x^T x^* = N$. Since $r \in \mathcal{S}$, we have $r = x^* \in \{0, 1\}^M$. This is in contradiction with the assumption $r \neq x^*$, hence $z^T E z > 0$, $\forall z \in Z_1$.

We next show that $Z_1$ and another set $Z_2$ defined by the following constraints are the same:

$$\left\| z \right\|_1 = f^T z = 1 \quad \text{(A.22)}$$

$$\sum_{i=1}^{M} z_i = 0 \quad \text{(A.23)}$$

$$z_i \in [0, 0.5] \quad \text{if } x_i^* = 0 \quad \text{(A.24)}$$

$$z_i \in [-0.5, 0] \quad \text{if } x_i^* = 1 \quad \text{(A.25)}$$

where $f \in \{-1, 1\}^M$ is as follows:

$$f_i = \begin{cases} 
1 & \text{if } x_i^* = 0 \\
-1 & \text{if } x_i^* = 1 
\end{cases} \quad \text{(A.26)}$$


- It’s easy to verify that if $z \in Z_1$, (A.22) and (A.23) hold. Since $r_i \in [0,1]$ and $x_i^* \in \{0,1\}$, if $x_i = 0$, $z_i \geq 0$; if $x_i = 1$, $z_i < 0$. On the other hand, if $|z_i| > 0.5$, then from (A.23) there must be another $|z_j| > 0.5$, $i \neq j$. This means that $\|z\|_1 \geq |z_i| + |z_j| > 1$, which contradicts (A.22). Hence $|z_i| \leq 0.5$, (A.24) and (A.25) hold. This proves that $z \in Z_2$.

- If $z \in Z_2$, there exist an $r = x^* + z$ that produces $\|z\|$ and it’s easy to verify that $r \in S$. Hence $z \in Z_1$.

$$\lambda(E) = \min_{z \in Z_2} z^T Ez > 0$$ is a convex problem, and can be computed efficiently using Quadratic programming.

\[\text{(A.27)}\]

\[= \eta^2 \|\nabla J(x_t)\|^2_2 - 2\eta \langle x_t - x^*, \nabla J(x_t) \rangle + (1 - \mu)\|x_t - x^*\|^2_2.\]

**Proof** In order to prove the global optimum convergence conditions of the projected gradient descent update given in (3.8), we first establish the global optimum convergence conditions of the gradient descent update.

**Step 1:** We first show the following (A.27) is less than 0:

\[
\|x_t - \eta \nabla J(x_t) - x^*\|^2_2 - \mu \|x_t - x^*\|^2_2 \leq \eta^2 \|\nabla J(x_t)\|^2_2 - 2\eta \langle x_t - x^*, \nabla J(x_t) \rangle + (1 - \mu)\|x_t - x^*\|^2_2.
\]

The gradient $\nabla J(x_t)$ is:

\[
\nabla J(x_t) = \sum_{y \in \mathcal{R}} A_y x_t \cdot (x_t^T A_y x_t - x^* A_y x^*)
\]

\[
= \sum_{y \in \mathcal{R}} A_y x_t \cdot (x_t - x^*)^T A_y (x_t + x^*).
\]

We then have:

\[
\|\nabla J(x_t)\|^2_2 = |\mathcal{R}|^2 \frac{1}{|\mathcal{R}|} \sum_{y} A_y x_t \cdot (x_t - x^*)^T A_y (x_t + x^*) \leq |\mathcal{R}|^2 \frac{1}{|\mathcal{R}|} \sum_{y} \|A_y x_t\|^2 \cdot (x_t - x^*)^T A_y (x_t + x^*) \leq |\mathcal{R}| \sum_{y} \|A_y x_t\|^2 \cdot ((x_t - x^*)^T A_y (x_t + x^*))^2 \]

\[
= 4|\mathcal{R}| \sigma_{\text{max}}^2 \|x_t\|^2_2 J(x_t),
\]

where $\sigma_{\text{max}}^2 \geq \sigma_{\text{max}}^2 (A_y), \forall y \in \mathcal{R}$; $\sigma_{\text{max}}^2$ is the largest squared maximum singular value of all $A_y$ and can be obtained using the Schur’s bound [59].
We also have:

\[
\langle x_t - x^*, \nabla J(x_t) \rangle = \sum_y (x_t - x^*)^T A_y x_t \cdot (x_t - x^*)^T A_y (x_t + x^*)
\]

\[
= 4J(x_t) \sum_y (x_t - x^*)^T A_y (x_t + x^*) \cdot (x_t - x^*)^T A_y x^* \tag{A.30}
\]

\[
\geq 4J(x_t) - \sqrt{4J(x_t)} \sqrt{\sum_y ((x_t - x^*)^T A_y x^*)^2}.
\]

Next we will lower-bound the above (A.30) further. Let \( h = x_t - x^* \), for some \( \frac{1}{2} < q < 1 \) we have:

\[
q^2 \sum_y ((x_t - x^*)^T A_y (x_t + x^*))^2 - \sum_y ((x_t - x^*)^T A_y x^*)^2
\]

\[
= q^2 \sum_y (h^T A_y (h + 2x^*))^2 - \sum_y (h^T A_y x^*)^2
\]

\[
= q^2 \sum_y (h^T A_y h)^2 + 4q^2 \sum_y h^T A_y h \cdot h^T A_y x^* + (4q^2 - 1) \sum_y (h^T A_y x^*)^2
\]

\[
\geq q^2 \sum_y (h^T A_y h)^2 - 4q^2 \sqrt{\sum_y (h^T A_y h)^2} \sqrt{\sum_y (h^T A_y x^*)^2} + (4q^2 - 1) \sum_y (h^T A_y x^*)^2
\]

\[
= q^2 \left( \sqrt{\sum_y (h^T A_y h)^2} - 2 \sqrt{\sum_y (h^T A_y x^*)^2} \right)^2 - \sum_y (h^T A_y x^*)^2
\]

\[
= \left( q \sqrt{\sum_y (h^T A_y h)^2} - (2q - 1) \sqrt{\sum_y (h^T A_y x^*)^2} \right) \left( q \sqrt{\sum_y (h^T A_y h)^2} - (2q + 1) \sqrt{\sum_y (h^T A_y x^*)^2} \right)
\]

\[
\tag{A.31}
\]

To make (A.31) greater than 0, either of the following two inequalities should hold:

\[
\sqrt{\sum_y (h^T A_y h)^2} > (2 + \frac{1}{q}) \sqrt{\sum_y (h^T A_y x^*)^2} \tag{A.32}
\]

\[
\sqrt{\sum_y (h^T A_y h)^2} < (2 - \frac{1}{q}) \sqrt{\sum_y (h^T A_y x^*)^2} \tag{A.33}
\]

We can obtain a lower bound on \( \tau = \|h\|_2 \) to make (A.33) hold. Specifically, the left-hand side of
(A.33) can be upper bounded via:

\[
\sum_y (h^T A_y h)^2 = \|h\|^4 \sum_y (u^T A_y u)^2
\]

\[
= \|h\|^4 \sum_y \|A_y\|^2_{op} \cdot \left( \frac{|u^T A_y u|}{\|A_y\|_{op}} \right)^2
\]

\[
\leq \|h\|^4 \sum_y \|A_y\|^2_{op} \cdot |u^T A_y u|
\]

\[
= \|h\|^4 \sum_y \sigma_{max}(A_y) \cdot |u^T A_y u|
\]

\[
\leq \|h\|^4 \cdot \sigma_{max} \cdot |u^T A| u
\]

\[
= \|h\|^2 \cdot \sigma_{max} \cdot |h^T A| h
\]

\[
= \|h\|^2 \cdot \sigma_{max} \cdot \|h\|^2_1
\]

where \( u = \frac{1}{\|h\|^2} h, z = \frac{1}{\|h\|^2} A_z, \sigma_{max} \) is the largest maximum singular value of all \( A_y \): \( \sigma_{max} \geq \sigma_{max}(A_y) \), and \( A = \sum_y A_y \) is a matrix of all 1s. The first inequality in (A.34) is obtained by \( |u^T A_y u| = |\langle u, A_y u \rangle| \leq \|u\|^2_2 |A_y u|_2 \leq \|A_y\|_{op} \) and hence \( \frac{|u^T A_y u|}{\|A_y\|_{op}} \leq 1 \); the second inequality is obtained by \( |u^T A_y u| = |\sum_{ij} A_y(i, j) u_i u_j| \leq \sum_{ij} A_y(i, j)|u_i||u_j| = |u^T A_y u| \). If we choose the operator norm \( \| \cdot \|_{op} \) to be the Euclidean norm, then \( \|A_y\|_{op} = \sigma_{max}(A_y) \).

The right-hand side of (A.33) can be low-bounded as:

\[
\sum_y (h^T A_y x)^2 = \|h\|^2_1 \cdot z^T \left( \sum_y A_y x^* x^* A_y^T \right) z = \|h\|^2_1 \cdot z^T E z
\]

\[
\geq \|h\|^2 \cdot \lambda(E),
\]

where \( E = \sum_{y \in \mathcal{R}} A_y x^* x^* A_y^T \). Combining (A.33), (A.34) and (A.35), we can see that as long as the following (A.36) holds, (A.33) will also hold.

\[
\|h\|_2 < \tau_{max} = \left( 2 - \frac{1}{q} \right) \cdot \sqrt{\frac{1}{\sigma_{max}} \lambda(E)}.
\]

(A.36) guarantees that (A.31) is always greater than 0. We then have \( -\sqrt{\sum_y ((x_t - x^*)^T A_y x^*)^2} > -q\sqrt{4J(x_t)} \). Plug this into (A.30), we have:

\[
\langle x_t - x^*, \nabla J(x_t) \rangle > 4(1 - q)J(x_t)
\]
Plug (A.29) and (A.37) into (A.27), we have:

\[
\left\| x_t - \eta \nabla J(x_t) - x^* \right\|^2_2 - \mu \left\| x_t - x^* \right\|^2_2 \\
< 4|R|\sigma_{\text{max}}^2 \left\| x_t \right\|^2_2 J(x_t) \cdot \eta^2 - 8(1 - q)J(x_t) \cdot \eta + (1 - \mu)\tau_{\text{max}}^2 \\
= 4|R|\sigma_{\text{max}}^2 \left\| x_t \right\|^2_2 J(x_t)
\times \left( \eta - \frac{(1 - q)}{|R|\sigma_{\text{max}}^2 \left\| x_t \right\|^2_2} \right)^2 - \frac{(1 - q)^2}{|R|^2\sigma_{\text{max}}^2 \left\| x_t \right\|^4_2} + \frac{(1 - \mu)\tau_{\text{max}}^2}{4|R|\sigma_{\text{max}}^2 \left\| x_t \right\|^2_2 J(x_t)} \tag{A.38}
\]

In order to make (A.38) strictly less than 0, the following should hold:

\[
\left( \eta - \frac{(1 - q)}{|R|\sigma_{\text{max}}^2 \left\| x_t \right\|^2_2} \right)^2 < \frac{(1 - q)^2}{|R|^2\sigma_{\text{max}}^2 \left\| x_t \right\|^4_2} - \frac{(1 - \mu)\tau_{\text{max}}^2}{4|R|\sigma_{\text{max}}^2 \left\| x_t \right\|^2_2 J(x_t)} \tag{A.39}
\]

The right hand side of (A.39) should be strictly greater than 0 so that a valid \( \eta \) can be obtained. This requires:

\[
\mu > 1 - \frac{(1 - q)^2}{|R|\sigma_{\text{max}}^2 \left\| x_t \right\|^2_2} \frac{4J(x_t)}{\tau_{\text{max}}^2} \tag{A.40}
\]

Also \( \mu \) should be greater than 0 to ensure the convergence. The step size \( \eta \) the gradient descent method should take is then:

\[
\frac{(1 - q)}{|R|\sigma_{\text{max}}^2 \left\| x_t \right\|^2_2} - \nu < \eta < \frac{(1 - q)}{|R|\sigma_{\text{max}}^2 \left\| x_t \right\|^2_2} + \nu, \tag{A.41}
\]

where \( \nu = \sqrt{\frac{(1 - q)^2}{|R|^2\sigma_{\text{max}}^2 \left\| x_t \right\|^4_2} - \frac{(1 - \mu)\tau_{\text{max}}^2}{4|R|\sigma_{\text{max}}^2 \left\| x_t \right\|^2_2 J(x_t)}}. \)

**Step 2:** Let \( x = x_t - \eta \nabla J(x_t) \) denote the gradient descent update, \( x_{t+1} = \mathcal{P}(x) \in S \) is the projected gradient descent update, \( s \) is a linear combination of \( x_{t+1} \) and \( x^* \) such that

\[
x^* - x_{t+1} = a(x_{t+1} - s), \tag{A.42}
\]

where \( a \in \mathbb{R}, a \neq 0 \) is some constant. If \( x \) is also a linear combination of \( x_{t+1} \) and \( x^* \), we can simply choose \( s = x \). If \( (x_{t+1} - x)^T(x^* - x_{t+1}) = 0 \), we can choose \( s = x_{t+1} \). Otherwise, we can still make the following hold by choosing

\[
a = \frac{|x^* - x_{t+1}|^2_2}{(x_{t+1} - x)^T(x^* - x_{t+1})} \\
(s - x)^T(s - x_{t+1}) = 0 \tag{A.43}
\]

In other words, we can always find an \( s \) that satisfies (A.43). From (A.43), we can get:

\[
\left\| x_t \right\|^2_2 - x_t^T x_t + \left\| x_{t+1} \right\|^2_2 - 2x_t^T x_{t+1} = -x_t^T x_{t+1} \tag{A.44}
\]

We also have:

\[
\left\| x - x_{t+1} \right\|^2_2 = \left\| x \right\|^2_2 + \left\| x_{t+1} \right\|^2_2 - 2x^T x_{t+1} \tag{A.45}
\]

\[
\left\| x - s \right\|^2_2 = \left\| x \right\|^2_2 + \left\| s \right\|^2_2 - 2x^T s \tag{A.46}
\]

\[
\left\| s - x_{t+1} \right\|^2_2 = \left\| s \right\|^2_2 + \left\| x_{t+1} \right\|^2_2 - 2s^T x_{t+1} \tag{A.47}
\]
Combining (A.44)-(A.47), we get:

$$\|x - x_{t+1}\|_2^2 = \|x - s\|_2^2 + \|s - x_{t+1}\|_2^2$$  \hspace{1cm} (A.48)

Similarly, we can get:

$$\|x - x^*\|_2^2 = \|x - s\|_2^2 + \|s - x^*\|_2^2$$  \hspace{1cm} (A.49)

• If $x \in S$, then $x = x_{t+1}$. (3.8) holds.

• If $x \notin S$ and $s \in S$, since $x_{t+1}$ is the projection of $x$ in $S$, we have $\|x - x_{t+1}\|_2 \leq \|x - s\|_2$. Combined with (A.48), we have:

$$\|s - x_{t+1}\|_2^2 = 0$$  \hspace{1cm} (A.50)

Hence $s$ and $x_{t+1}$ is the same point. From (A.49), we can get:

$$\|x - x^*\|_2^2 = \|x - x_{t+1}\|_2^2 + \|x_{t+1} - x^*\|_2^2$$  \hspace{1cm} (A.51)

Since $x \notin S$, we have $\|x - x_{t+1}\|_2^2 > 0$. Hence $\|x - x^*\|_2^2 > \|x_{t+1} - x^*\|_2^2$. (3.8) holds.

• If $x \notin S$ and $s \notin S$, we have:

$$\|s - x^*\|_2^2 = \|s - x_{t+1} + x_{t+1} - x^*\|_2^2$$

$$= \|s - x_{t+1}\|_2^2 + \|x_{t+1} - x^*\|_2^2 + 2 (s - x_{t+1})^T (x_{t+1} - x^*)$$  \hspace{1cm} (A.52)

- If $a \in [0, \infty)$, from (A.42), we have $(s - x_{t+1})^T (x_{t+1} - x^*) \geq 0$. From (A.52), we have $\|s - x^*\|_2^2 \geq \|s - x_{t+1}\|_2^2$. Combined with (A.48) and (A.49), we have $\|x - x^*\|_2^2 \geq \|x_{t+1} - x^*\|_2^2$. (3.8) holds.

- If $a \in (-1, 0)$, from (A.42), we have $x_{t+1} - x^* = \frac{a}{1-a} (x^* - s)$. Since $\frac{a}{1-a} > 0$, $(x_{t+1} - x^*)^T (x^* - s) > 0$. We then have:

$$\|x_{t+1} - s\|_2^2 = \|x_{t+1} - x^* + x^* - s\|_2^2$$

$$= \|x_{t+1} - x^*\|_2^2 + \|x^* - s\|_2^2 + 2 (x_{t+1} - x^*)^T (x^* - s)$$ \hspace{1cm} (A.53)

Combined with (A.48) and (A.49), we have:

$$\|x - x_{t+1}\|_2^2 > \|x - x^*\|_2^2$$  \hspace{1cm} (A.54)

This is in contradiction with the fact that $x_{t+1}$ is the projection of $x$ in $S$ so that $\|x - x_{t+1}\|_2^2$ has the smallest $l_2$ norm: $\|x - x_{t+1}\|_2^2 \leq \|x - x^*\|_2^2$, hence $a \notin (-1, 0)$.

- If $a \in (-\infty, -1]$, from (A.42), we have $s = -\frac{1}{a} x^* + (1 + \frac{1}{a}) x_{t+1}$. Since $-\frac{1}{a} \notin (0, 1], s \in S$. This is in contradiction with the assumption $s \notin S$, hence $a \notin (-\infty, 1]$.

In summary, we now have:

$$\|x_{t+1} - x^*\|_2^2 \leq \|x_t - \eta \cdot \nabla J(x_t) - x^*\|_2^2 < \mu \|x_t - x^*\|_2^2.$$  \hspace{1cm} (A.55)
A.7 Theorem 3.3

Proof This is also a 2-steps procedure as in Proof A.6 of Theorem 3.2. Let $\mathcal{R}' = \{ y \mid F(x) > 0 \} \subset \mathcal{R}$, we first establish the global optimum convergence conditions of the gradient descent update.

Step 1: We first show the following \((A.56)\) is less than 0:

\[
\|x_t - \eta \nabla J(x_t) - x^*\|_2^2 - \mu \|x_t - x^*\|_2^2 = \eta^2\|\nabla J(x_t)\|_2^2 - 2\eta \langle x_t - x^*, \nabla J(x_t) \rangle + (1 - \mu)\|x_t - x^*\|_2^2 .
\]

(A.56)

The gradient $\nabla J(x_t)$ is:

\[
\nabla J(x_t) = \sum_{y \in \mathcal{R}'} -A_y x_t \cdot \sqrt{\frac{F(y)}{x_t^T A_y x_t}}.
\]

We then have:

\[
\|\nabla J(x_t)\|_2^2 = |\mathcal{R}'|^2 \left[\frac{1}{|\mathcal{R}'|} \sum_y \frac{-A_y x_t \cdot \sqrt{F(y)}}{x_t^T A_y x_t}\right]^2
\]

\[
\leq |\mathcal{R}'|^2 \frac{1}{|\mathcal{R}'|} \sum_y \|A_y x_t\|_2 \frac{F(y)}{x_t^T A_y x_t}
\]

\[
\leq |\mathcal{R}'| \sigma_{\text{max}}^2 \|x_t\|_2 \sum_y \frac{F(y)}{x_t^T A_y x_t}.
\]

(A.58)

where $\sigma_{\text{max}}^2 \geq \sigma_{\text{max}}^2(A_y), \forall y \in \mathcal{R}'$.

We further have:

\[
\langle x_t - x^*, \nabla J(x_t) \rangle = \sum_y (x_t - x^*)^T A_y x_t \cdot -\sqrt{\frac{F(y)}{x_t^T A_y x_t}}
\]

\[
= \sum_y (x_t - x^*)^T A_y x_t \cdot \left(1 - \sqrt{\frac{F(y)}{x_t^T A_y x_t}}\right)
\]

\[
= \sum_y (x_t - x^*)^T A_y x_t \cdot \frac{(x_t - x^*)^T A_y (x_t + x^*)}{\sqrt{x_t^T A_y x_t} \cdot \left(\sqrt{x_t^T A_y x_t + \sqrt{F(y)}}\right)}
\]

\[
= \sum_y (x_t - x^*)^T A_y x_t \cdot \frac{(x_t - x^*)^T A_y (x_t + x^*)}{H(y)}
\]

\[
= \sum_y \left(\frac{(x_t - x^*)^T A_y (x_t + x^*)}{H(y)}\right)^2 - \sum_y \left(\frac{(x_t - x^*)^T A_y (x_t + x^*) \cdot (x_t - x^*)^T A_y x^*}{H(y)}\right)
\]

(A.59)
where \( H(y) = \sqrt{x_t^T A_y x_t \cdot \left( x_t^T A_y x_t + \sqrt{F(y)} \right)} \). Let \( h = x_t - x^* \), for some \( \frac{1}{2} < q < 1 \), we can lower-bound the above (A.59) similarly as is done in (A.30). We have:

\[
q^2 \sum_y \frac{((x_t - x^*)^T A_y (x_t + x^*))^2}{H(y)} - \sum_y \frac{((x_t - x^*)^T A_y x^*)^2}{H(y)}.
\]  

(A.60)

We need the following inequalities to hold to make (A.60) greater than 0.

\[
\sqrt{\sum_y \frac{(h^T A_y h)^2}{H(y)}} > (2 + \frac{1}{q}) \sqrt{\sum_y \frac{(h^T A_y x^*)^2}{H(y)}}
\]  

(A.61)

\[
\sqrt{\sum_y \frac{(h^T A_y h)^2}{H(y)}} < (2 - \frac{1}{q}) \sqrt{\sum_y \frac{(h^T A_y x^*)^2}{H(y)}}
\]  

(A.62)

We can also obtain a lower bound on \( \tau = \|h\|_2 \) to make (A.62) hold. We first establish a lower bound for \( H(y) \). Since we assume \( \|h\|_2 = \sqrt{\sum_i ((x_i)_t - x^*_i)^2} \leq \phi < x^*_i \), where \( x^*_i \) is the minimum nonzero entry of \( x^* \), we have:

\[
|(x_i)_t - x^*_i| \leq \phi, \quad \forall i \in [M].
\]  

(A.63)

We then have \( (x_i)_t \geq x^*_i - \phi > 0 \). Hence, \( \sqrt{x_t^T A_y x_t} \geq \min_{ij} \sqrt{(x_i)_t \cdot (x_j)_t} \geq x^*_i - \phi \). Let \( F(y) \geq F_{\min}, \forall y \in \mathcal{R}' \), \( H(y) \) can be lower-bounded as follows:

\[
H(y) \geq (x^*_i - \phi) \cdot (x^*_i - \phi + \sqrt{F_{\min}}) = \kappa_{\min} > 0.
\]  

(A.64)

Let \( F(y) \leq F_{\max} = N, \forall y \in \mathcal{R}' \), since \( 0 \leq (x_i)_t \leq 1 \) and \( \|x_t\|_1 = N, H(y) \) can be upper-bounded as follows:

\[
H(y) \leq \sqrt{N} (\sqrt{N} + \sqrt{N}) = 2N = \kappa_{\max}.
\]  

(A.65)

Similarly to what is done in (A.34), the left-hand side of (A.62) can be upper bounded:

\[
\sum_y \frac{(h^T A_y h)^2}{H(y)} \leq \frac{1}{\kappa_{\min}} \cdot \|h\|_2^2 \cdot \sigma_{\max} \cdot \|h\|_1^2.
\]  

(A.66)

The right-hand side of (A.62) can be lower bounded as:

\[
\sum_y \frac{(h^T A_y x^*)^2}{H(y)} \geq \frac{1}{\kappa_{\max}} \cdot \|h\|_1^2 \cdot \lambda(E),
\]  

(A.67)

where \( E = \sum_{y \in \mathcal{R}'} A_y x^* x^* A_y^T \). Combining (A.62), (A.66) and (A.67), we can see as long as the following (A.68) holds, (A.62) will hold:

\[
\|h\|_2 < \tau_{\max} = \min \left\{ \phi, \left( 2 - \frac{1}{q} \right) \cdot \sqrt{\frac{1}{\sigma_{\max} \kappa_{\max}} \lambda(E)} \right\}.
\]  

(A.68)
Plug (A.60) > 0 to (A.59), we then have:

\[
\langle x_t - x^*, \nabla J(x_t) \rangle > (1 - q) \sum_y \frac{((x_t - x^*)^T A_y (x_t + x^*))^2}{H(y)}
\]  

(A.69)

Plug (A.58) and (A.69) into (A.56), we have:

\[
\| x_t - \eta \nabla J(x_t) - x^* \|^2 - \mu \| x_t - x^* \|^2
\]

\[
= \frac{\lambda_{R'}^2}{\sigma_{y, \text{max}}^2} \sum_y \frac{F(y)}{x_t^T A_y x_t} \cdot \eta^2 - 2(1 - q) \sum_y \frac{((x_t - x^*)^T A_y (x_t + x^*))^2}{H(y)} \cdot \eta + (1 - \mu) \tau_{\text{max}}^2
\]

\[
= \frac{\lambda_{R'}^2}{\sigma_{y, \text{max}}^2} \sum_y \frac{F(y)}{x_t^T A_y x_t} \cdot \eta^2 \left( \eta - \frac{(1 - q) \Gamma_2}{|\lambda_{R'}^2 \sigma_{y, \text{max}}^2 \| x_t \|_2^2 \Gamma_1} \right)^2 - \frac{(1 - q)^2 \Gamma_2^2}{|\lambda_{R'}^2 \sigma_{y, \text{max}}^2 \| x_t \|_2^2 \Gamma_1^2} + \frac{(1 - \mu) \tau_{\text{max}}^2}{|\lambda_{R'}^2 \sigma_{y, \text{max}}^2 \| x_t \|_2^2 \Gamma_1}
\]  

(A.70)

where \( \Gamma_1 = \sum_y \frac{F(y)}{x_t^T A_y x_t} \), \( \Gamma_2 = \sum_y \frac{((x_t - x^*)^T A_y (x_t + x^*))^2}{H(y)} \).

In order to make (A.70) strictly less than 0, the following should hold:

\[
\left( \eta - \frac{(1 - q) \Gamma_2}{|\lambda_{R'}^2 \sigma_{y, \text{max}}^2 \| x_t \|_2^2 \Gamma_1} \right)^2 < \frac{(1 - q)^2 \Gamma_2^2}{|\lambda_{R'}^2 \sigma_{y, \text{max}}^2 \| x_t \|_2^2 \Gamma_1^2} - \frac{(1 - \mu) \tau_{\text{max}}^2}{|\lambda_{R'}^2 \sigma_{y, \text{max}}^2 \| x_t \|_2^2 \Gamma_1}
\]  

(A.71)

The right hand side of (A.71) should be strictly greater than 0 so that a valid \( \eta \) can be obtained. This requires:

\[
\mu > 1 - \frac{(1 - q)^2 \Gamma_2^2}{|\lambda_{R'}^2 \sigma_{y, \text{max}}^2 \| x_t \|_2^2 \Gamma_1 \tau_{\text{max}}^2}
\]  

(A.72)

Also \( \mu \) should be greater than 0 to ensure the convergence. The step size \( \eta \) the gradient descent method should take is then:

\[
\frac{(1 - q) \Gamma_2}{|\lambda_{R'}^2 \sigma_{y, \text{max}}^2 \| x_t \|_2^2 \Gamma_1} - \nu < \eta < \frac{(1 - q) \Gamma_2}{|\lambda_{R'}^2 \sigma_{y, \text{max}}^2 \| x_t \|_2^2 \Gamma_1} + \nu,
\]  

(A.73)

where \( \nu = \sqrt{\frac{(1 - q)^2 \Gamma_2^2}{|\lambda_{R'}^2 \sigma_{y, \text{max}}^2 \| x_t \|_2^2 \Gamma_1^2} - \frac{(1 - \mu) \tau_{\text{max}}^2}{|\lambda_{R'}^2 \sigma_{y, \text{max}}^2 \| x_t \|_2^2 \Gamma_1}} \).

Step 2: The proof of the global convergence conditions of the projected gradient descent update is the same as the step-2 proof A.6 of Theorem 3.2.

A.8 Theorem 3.4

Proof This is also a 2-steps procedure as in Proof A.6 of Theorem 3.2. Let \( R' = \{ y \mid F(y) > 0 \} \subset R \), we first establish the global optimum convergence conditions of the gradient descent update. The following is an adaptation of Proof A.7.
Step 1: We first show the following (A.74) is less than 0:

\[ \| x_t - \eta \nabla J(x_t) - x^* \|^2_2 - \mu \| x_t - x^* \|^2_2 \]
\[ = \eta^2 \| \nabla J(x_t) \|^2_2 - 2\eta \langle x_t - x^*, \nabla J(x_t) \rangle + (1 - \mu) \| x_t - x^* \|^2_2. \] (A.74)

The gradient \( \nabla J(x_t) \) is:

\[ \nabla J(x_t) = \sum_{y \in R'} -A_y x_t \cdot \frac{F(y)}{x_t^T A_y x_t}. \] (A.75)

We then have:

\[ \| \nabla J(x_t) \|^2_2 = \| R' \| \frac{1}{| R' |} \sum_{y} -A_y x_t \cdot \frac{F(y)}{x_t^T A_y x_t} \|^2_2 \]
\[ \leq \| R' \| \frac{1}{| R' |} \sum_{y} \| A_y x_t \|^2_2 \left( \frac{F(y)}{x_t^T A_y x_t} \right)^2 \] (A.76)
\[ \leq | R' | \sigma_{\max}^2 \| x_t \|^2_2 \sum_{y} \left( \frac{F(y)}{x_t^T A_y x_t} \right)^2. \]

where \( \sigma_{\max}^2 \geq \sigma_{\max}^2 (A_y), \forall y \in \mathcal{R}'. \)

We further have:

\[ \langle x_t - x^*, \nabla J(x_t) \rangle = \sum_{y} (x_t - x^*)^T A_y x_t \cdot \left( 1 - \frac{F(y)}{x_t^T A_y x_t} \right) \]
\[ = \sum_{y} (x_t - x^*)^T A_y x_t \cdot \frac{(x_t - x^*)^T A_y (x_t + x^*)}{x_t^T A_y x_t} \]
\[ = \sum_{y} \left( \frac{(x_t - x^*)^T A_y (x_t + x^*)}{x_t^T A_y x_t} \right)^2 - \sum_{y} \left( \frac{(x_t - x^*)^T A_y (x_t + x^*) \cdot (x_t - x^*)^T A_y x_t}{x_t^T A_y x_t} \right) \]
\[ \geq \sum_{y} \left( \frac{(x_t - x^*)^T A_y (x_t + x^*)}{x_t^T A_y x_t} \right)^2 - \sum_{y} \left( \frac{(x_t - x^*)^T A_y (x_t + x^*)}{x_t^T A_y x_t} \right) \]
\[ = q^2 \sum_{y} \left( \frac{(x_t - x^*)^T A_y (x_t + x^*)}{x_t^T A_y x_t} \right)^2 - \sum_{y} \left( \frac{(x_t - x^*)^T A_y x^*}{x_t^T A_y x_t} \right)^2. \] (A.77)

Let \( h = x_t - x^* \), for some \( \frac{1}{2} < q < 1 \), we can lower-bound the above (A.77) similarly as is done in (A.30). We have:

\[ q^2 \sum_{y} \left( \frac{(x_t - x^*)^T A_y (x_t + x^*)}{x_t^T A_y x_t} \right)^2 - \sum_{y} \left( \frac{(x_t - x^*)^T A_y x^*}{x_t^T A_y x_t} \right)^2. \] (A.78)
We need the following inequalities to hold to make (A.78) greater than 0.

\[
\begin{align*}
&\sum_y \frac{(h^T A_y h)^2}{x_t^T A_y x_t} > (2 + \frac{1}{q}) \sum_y \frac{(h^T A_y x^*)^2}{x_t^T A_y x_t} \\
&\sum_y \frac{(h^T A_y h)^2}{x_t^T A_y x_t} < (2 - \frac{1}{q}) \sum_y \frac{(h^T A_y x^*)^2}{x_t^T A_y x_t}
\end{align*}
\]  \quad (A.79) \quad (A.80)

We can also obtain a lower bound on \( \tau = \|h\|_2 \) to make (A.80) hold. We first establish a lower bound for \( x_t^T A_y x_t \). Since we assume \( \|h\|_2 = \sqrt{\sum_i ((x_t)_i - x^*_i)^2} < \phi < x^*_\min \), where \( x^*_\min \) is the minimum nonzero entry of \( x^* \), we have:

\[
|(x_t)_i - x^*_i| \leq \phi, \quad \forall i \in [M].
\]  \quad (A.81)

We then have \( (x_t)_i \geq x^*_\min - \phi > 0 \). \( x_t^T A_y x_t \) can be lower-bounded as follows:

\[
x_t^T A_y x_t \geq (x^*_\min - \phi)^2 = \kappa_{\min} > 0.
\]  \quad (A.82)

Since \( 0 \leq (x_t)_i \leq 1 \) and \( \|x_t\|_1 = N \), \( x_t^T A_y x_t \) can be upper-bounded as follows:

\[
x_t^T A_y x_t \leq N = \kappa_{\max}.
\]

Similarly to what is done in (A.34), the left-hand side of (A.80) can be upper bounded:

\[
\sum_y \frac{(h^T A_y h)^2}{x_t^T A_y x_t} \leq \frac{1}{\kappa_{\min}} \cdot \|h\|_2^2 \cdot \sigma_{\max} \cdot \|h\|_1^2.
\]  \quad (A.84)

The right-hand side of (A.80) can be lower bounded as:

\[
\sum_y \frac{(h^T A_y x^*)^2}{x_t^T A_y x_t} \geq \frac{1}{\kappa_{\max}} \cdot \|h\|_1^2 \cdot \lambda(E),
\]  \quad (A.85)

where \( E = \sum_{y \in \mathcal{R}^c} A_y x^* x^{*T} A_y^T \). Combining (A.80), (A.84) and (A.85), we can see as long as the following (A.86) holds, (A.80) will hold:

\[
\|h\|_2 < \tau_{\max} = \min \left\{ \phi, \left( 2 - \frac{1}{q} \right) \frac{1}{\sigma_{\max}} \kappa_{\min} \lambda(E) \right\}.
\]  \quad (A.86)

Plug (A.78) > 0 to (A.77), we then have:

\[
\langle x_t - x^*, \nabla J(x_t) \rangle > (1 - q) \sum_y \frac{((x_t - x^*)^T A_y (x_t + x^*))^2}{x_t^T A_y x_t}
\]  \quad (A.87)

Plug (A.76) and (A.87) into (A.74), we have:

\[
\|x_t - \eta \nabla J(x_t) - x^*\|_2^2 - \mu \|x_t - x^*\|_2^2
\]

\[
= |\mathcal{R}'| \sigma_{\max}^2 \|x_t\|_2^2 \sum_y F(y) x_t^T A_y x_t \cdot \eta^2 - 2(1 - q) \sum_y \frac{((x_t - x^*)^T A_y (x_t + x^*))^2}{x_t^T A_y x_t} \cdot \eta + (1 - \mu) \tau_{\max}^2
\]

\[
= |\mathcal{R}'| \sigma_{\max}^2 \|x_t\|_2^2 \Gamma_1 \cdot \left( \eta - \frac{(1 - q) \Gamma_2}{|\mathcal{R}'| \sigma_{\max}^2 \|x_t\|_2^2 \Gamma_1} \right)^2 - \frac{(1 - q)^2 \Gamma_2^2}{|\mathcal{R}'| \sigma_{\max}^2 \|x_t\|_2^2 \Gamma_1} + \frac{(1 - \mu) \tau_{\max}^2}{|\mathcal{R}'| \sigma_{\max}^2 \|x_t\|_2^2 \Gamma_1},
\]  \quad (A.88)
where \( \Gamma_1 = \sum_y f(y) A_y \), \( \Gamma_2 = \sum_y \frac{(x_1 - x^*)^T A_y (x_1 + x^*))^2}{x_1^T A_y x_1} \).

In order to make (A.88) strictly less than 0, the following should hold:

\[
\left( \eta - \frac{(1-q)\Gamma_2}{|\mathcal{R}'|\sigma^2_{\text{max}} \|x_t\|^2 \Gamma_1} \right)^2 < \frac{(1-q)^2 \Gamma_2^2}{|\mathcal{R}'|^2 \sigma^4_{\text{max}} \|x_t\|^4 \Gamma_1^2} - \frac{(1-\mu)^2 \Gamma_2^2}{|\mathcal{R}'|^2 \sigma^4_{\text{max}} \|x_t\|^4 \Gamma_1^2} \tag{A.89}
\]

The right hand side of (A.89) should be strictly greater than 0 so that a valid \( \eta \) can be obtained. This requires:

\[
\mu > 1 - \frac{(1-q)^2 \Gamma_2^2}{|\mathcal{R}'|^2 \sigma^4_{\text{max}} \|x_t\|^4 \Gamma_1^2} \tag{A.90}
\]

Also \( \mu \) should be greater than 0 to ensure the convergence. The step size \( \eta \) the gradient descent method should take is then:

\[
\frac{(1-q)\Gamma_2}{|\mathcal{R}'|\sigma^2_{\text{max}} \|x_t\|^2 \Gamma_1} - \nu < \eta < \frac{(1-q)\Gamma_2}{|\mathcal{R}'|\sigma^2_{\text{max}} \|x_t\|^2 \Gamma_1} + \nu, \tag{A.91}
\]

where \( \nu = \sqrt{\frac{(1-q)^2 \Gamma_2^2}{|\mathcal{R}'|\sigma^4_{\text{max}} \|x_t\|^4 \Gamma_1^2} - \frac{(1-\mu)^2 \Gamma_2^2}{|\mathcal{R}'|\sigma^4_{\text{max}} \|x_t\|^4 \Gamma_1^2}} \).

**Step 2:** The proof of the global convergence conditions of the projected gradient descent update is the same as the step-2 proof A.6 of Theorem 3.2.

---

**References**

[1] M. M. Abbas and H. M. Bahig, “A fast exact sequential algorithm for the partial digest problem,” BMC Bioinformatics, vol. 17, no. 19, p. 510, Dec 2016.

[2] H. Ahrabian, M. Ganjtabesh, A. Nowzari-Dalini, and Z. Razaghi-Moghadam-Kashani, “Genetic algorithm solution for partial digest problem,” Int. J. Bioinformatics Res. Appl., vol. 9, no. 6, pp. 584–594, Sep 2013.

[3] S. M. Ali and S. D. Silvey, “A general class of coefficients of divergence of one distribution from another,” Journal of the Royal Statistical Society. Series B (Methodological), vol. 28, no. 1, pp. 131–142, 1966.

[4] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating smallroom acoustics,” The Journal of the Acoustical Society of America, vol. 65, no. 4, pp. 943–950, 1979.

[5] N. K. Batmanghelich, B. Taskar, and C. Davatzikos, “Generative-discriminative basis learning for medical imaging,” IEEE Transactions on Medical Imaging, vol. 31, no. 1, pp. 51–69, Jan 2012.

[6] A. Bhattacharyya, “On a measure of divergence between two statistical populations defined by their probability distributions,” Bulletin of the Calcutta Mathematical Society, vol. 35, pp. 99–109, 1943.
[7] S. J. L. Billinge, “The atomic pair distribution function: past and present,” Zeitschrift fr Kristallographie - Crystalline Materials, vol. 219, no. 3, pp. 117–121, 2004.

[8] ——, “Viewpoint: The nanostructure problem,” Physics, vol. 3, no. 25, 2010.

[9] S. J. L. Billinge, P. M. Duxbury, D. S. Gonçalves, C. Lavor, and A. Mucherino, “Assigned and unassigned distance geometry: applications to biological molecules and nanostructures,” 4OR, vol. 14, no. 4, pp. 337–376, Dec 2016.

[10] S. J. L. Billinge, P. M. Duxbury, D. S. Gonçalves, C. Lavor, and A. Mucherino, “Assigned and unassigned distance geometry: applications to biological molecules and nanostructures,” 4OR, vol. 14, no. 4, pp. 337–376, Apr. 2016.

[11] S. J. L. Billinge and I. Levin, “The problem with determining atomic structure at the nanoscale,” Science, vol. 316, no. 5824, pp. 561–565, 2007.

[12] J. Blazewicz, P. Formanowicz, M. Kasprzak, M. Jaroszewski, and W. Markiewicz, “Construction of dna restriction maps based on a simplified experiment,” Bioinformatics, vol. 17, no. 5, pp. 398–404, May 2001.

[13] J. Blazewicz, E. Burke, M. Kasprzak, A. Kovalev, and M. Kovalyov, “Simplified partial digest problem: Enumerative and dynamic programming algorithms,” IEEE/ACM Trans. Comput. Biol. Bioinformatics, vol. 4, no. 4, pp. 668–680, Oct. 2007.

[14] A. A. Bothner-By, R. L. Stephens, J. Lee, C. D. Warren, and R. W. Jeanloz, “Structure determination of a tetrasaccharide: transient nuclear overhauser effects in the rotating frame,” Journal of the American Chemical Society, vol. 106, no. 3, pp. 811–813, 1984.

[15] M. Boutin and G. Kemper, “On Reconstructing N-Point Configurations From the Distribution of Distances or Areas,” in Algorithms in Bioinformatics. Springer, Berlin, Heidelberg, Sep. 2003, pp. 111–123.

[16] G. Buchanan, “Nuclear magnetic resonance studies of crown ethers,” Progress in Nuclear Magnetic Resonance Spectroscopy, vol. 34, no. 3, pp. 327 – 377, 1999.

[17] M. Cieliebak, S. Eidenbenz, and P. Penna, “Noisy Data Make the Partial Digest Problem NP-hard,” in Algorithms in Bioinformatics. Springer, Berlin, Heidelberg, Sep. 2003, pp. 111–123.

[18] T. M. Cover and J. A. Thomas, Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, 2006.

[19] M. Crocco, A. Trucco, V. Murino, and A. Del Bue, “Towards Fully Uncalibrated Room Reconstruction with Sound,” in EUSIPCO. Lisabon: IEEE, 2014, pp. 910–914.

[20] I. Csiszár, “Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von markoffschen ketten,” Magyar. Tud. Akad. Mat. Kutató Int. Közl, vol. 8, pp. 85–108, 1963.

[21] T. Dakic, “On the turnpike problem,” Ph.D. dissertation, Simon Fraser University, 2000.
[22] K. J. Danna, “Determination of fragment order through partial digests and multiple enzyme digests,” in Nucleic Acids Part I, ser. Methods in Enzymology. Academic Press, 1980, vol. 65, no. Supplement C, pp. 449 – 467.

[23] I. Dokmanić, L. Daudet, and M. Vetterli, “How to Localize Ten Microphones in One Fingersnap,” in EUSIPCO, 2014.

[24] ——, “From acoustic room reconstruction to SLAM,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 6345–6349.

[25] I. Dokmanić, Y. M. Lu, and M. Vetterli, “Can One Hear the Shape of a Room: The 2-D Polygonal Case,” in IEEE ICASSP. Prague: IEEE, 2011, pp. 321–324.

[26] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, “Acoustic Echoes Reveal Room Shape,” Proc. Natl. Acad. Sci., vol. 110, no. 30, pp. 12 186–12 191, Jun. 2013.

[27] P. M. Duxbury and S. J. L. Billinge, “Graph rigidity, unassigned distance geometry and the nanostructure problem,” in 50th Asilomar Conference on Signals, Systems and Computers, Nov 2016, pp. 1483–1487.

[28] P. Duxbury, L. Granlund, S. Gujarathi, P. Juhas, and S. Billinge, “The unassigned distance geometry problem,” Discrete Applied Mathematics, vol. 204, no. Supplement C, pp. 117 – 132, 2016.

[29] S. R. Gujarathi, C. L. Farrow, C. Glosser, L. Granlund, and P. M. Duxbury, “Ab-Initio Reconstruction of Complex Euclidean Networks in Two Dimensions,” Physical Review E, vol. 89, no. 5, 2014.

[30] M. D. Gupta, S. Kumar, and J. Xiao, “L1 projections with box constraints,” CoRR, vol. abs/1010.0141, 2010.

[31] E. Hellinger, “Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen.” Journal für die reine und angewandte Mathematik, vol. 136, pp. 210–271, 1909.

[32] J. D. Hoheisel, D. Nizetic, and H. Lehrach, “Control of partial digestion combining the enzymes dam methylase and mboi.” Nucleic Acids Research, vol. 17, no. 23, pp. 9571–9582, 1989.

[33] I. Jager, R. Heusdens, and N. D. Gauibitch, “Room geometry estimation from acoustic echoes using graph-based echo labeling,” in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 1–5.

[34] C. Johnson, “Diffusion ordered nuclear magnetic resonance spectroscopy: principles and applications,” Progress in Nuclear Magnetic Resonance Spectroscopy, vol. 34, no. 3, pp. 203–256, 1999.

[35] P. Juhas, D. M. Cherba, P. M. Duxbury, W. F. Punch, and S. J. L. Billinge, “Ab initio determination of solid-state nanostructure,” Nature, vol. 440, no. 7084, pp. 655–658, Mar. 2006.
[36] P. Juhas, L. Granlund, P. M. Duxbury, W. F. Punch, S. J. L. Billinge, and IUCr, “The Liga algorithm for ab initio determination of nanostructure,” Acta Crystallogr Sect A Found Crystallogr, vol. 64, no. 6, pp. 631–640, Nov. 2008.

[37] J. W. Keepers and T. L. James, “A theoretical study of distance determinations from nmr. two-dimensional nuclear overhauser effect spectra,” Journal of Magnetic Resonance (1969), vol. 57, no. 3, pp. 404 – 426, 1984.

[38] P. Koehl, “Linear prediction spectral analysis of nmr data,” Progress in Nuclear Magnetic Resonance Spectroscopy, vol. 34, p. 257299, 05 1999.

[39] H. W. Kuhn, “The hungarian method for the assignment problem,” Naval Research Logistics Quarterly, vol. 2, no. 12, pp. 83–97, 1955.

[40] A. Kumar, R. Ernst, and K. Wthrich, “A two-dimensional nuclear overhauser enhancement (2d noe) experiment for the elucidation of complete proton-proton cross-relaxation networks in biological macromolecules,” Biochemical and Biophysical Research Communications, vol. 95, no. 1, pp. 1 – 6, 1980.

[41] P. Lemke, S. S. Skiena, and W. D. Smith, “Reconstructing Sets From Interpoint Distances,” in Discrete and Computational Geometry. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 597–631.

[42] E. Levina and P. Bickel, “The earth mover’s distance is the mallows distance: some insights from statistics,” in Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001, vol. 2, 2001, pp. 251–256 vol.2.

[43] C. L. Mallows, “A note on asymptotic joint normality,” The Annals of Mathematical Statistics, vol. 43, no. 2, pp. 508–515, 1972.

[44] D. Marković, F. Antonacci, A. Sarti, and S. Tubaro, “Estimation of Room Dimensions From a Single Impulse Response,” IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 1–4, 2013.

[45] H. Mohimani, “Bioinformatics methods for natural product discovery,” Ph.D. dissertation, UCSD, 2013.

[46] H. Mohimani, A. Gurevich, A. Mikheenko, N. Garg, L.-F. Nothias, A. Ninomiya, K. Takada, P. C. Dorrestein, and P. A. Pevzner, “Dereplication of peptidic natural products through database search of mass spectra,” Nature Chemical Biology, vol. 13, no. 1, pp. 30–37, Jan. 2017.

[47] H. Mohimani, W.-T. Liu, Y.-L. Yang, S. P. Gaudêncio, W. Fenical, P. C. Dorrestein, and P. A. Pevzner, “Multiplex De Novo Sequencing of Peptide Antibiotics,” Journal of Computational Biology, vol. 18, no. 11, pp. 1371–1381, Nov. 2011.

[48] A. H. Moore, M. Brookes, and P. A. Naylor, “Room Geometry Estimation From a Single Channel Acoustic Impulse Response,” in EUSIPCO, 2013, pp. 1–5.

[49] J. Morales, J. Patton, and B. J.W., “Partial endonuclease digestion mapping of restriction sites using pcr-amplified dna.” Genome research, vol. 2, pp. 228–233, 1993.
[50] T. Morimoto, “Markov processes and the h-theorem,” *Journal of the Physical Society of Japan*, vol. 18, no. 3, pp. 328–331, 1963.

[51] R. Nadimi, H. S. Fathabadi, and M. Ganjtabesh, “A fast algorithm for the partial digest problem,” *Japan Journal of Industrial and Applied Mathematics*, vol. 28, no. 2, p. 315, Aug 2011.

[52] D. Neuhaus, *Nuclear Overhauser Effect*. American Cancer Society, 2011.

[53] J. H. NOGGLE and R. E. SCHIRMER, Eds., *Front Matter*. Academic Press, 1971.

[54] V. Petkov, *Pair Distribution Functions Analysis*. American Cancer Society, 2012, pp. 1–14.

[55] V. Petkov, B. Prasai, Y. Ren, S. Shan, J. Luo, P. Joseph, and C.-J. Zhong, “Solving the nanostructure problem: exemplified on metallic alloy nanoparticles,” *Nanoscale*, vol. 6, no. 17, pp. 10 048–10 061, Sep 2014.

[56] P. A. Pevzner, “Dna physical mapping and alternating eulerian cycles in colored graphs,” *Algorithmica*, vol. 13, no. 1, pp. 77–105, Feb 1995.

[57] F. Ribeiro, D. A. Florencio, D. E. Ba, and C. Zhang, “Geometrically Constrained Room Modeling With Compact Microphone Arrays,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 20, no. 5, pp. 1449–1460, 2012.

[58] J. Rosenblatt and P. D. Seymour, “The structure of homometric sets,” *SIAM Journal on Algebraic Discrete Methods*, vol. 3, no. 3, pp. 343–350, 1982.

[59] J. Schur, “Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen.” *Journal für die reine und angewandte Mathematik*, vol. 140, pp. 1–28, 1911.

[60] S. Shalev-Shwartz and Y. Singer, “Efficient learning of label ranking by soft projections onto polyhedra,” *J. Mach. Learn. Res.*, vol. 7, pp. 1567–1599, Dec. 2006.

[61] S. S. Skiena, W. D. Smith, and P. Lemke, “Reconstructing sets from interpoint distances (extended abstract),” in *Proceedings of the Sixth Annual Symposium on Computational Geometry*. ACM, 1990, pp. 332–339.

[62] S. S. Skiena and G. Sundaram, “A partial digest approach to restriction site mapping,” *Bull. Math. Biol.*, vol. 56, no. 2, pp. 275–294, 1994.

[63] S. Williams, “Cerebral amino acids studied by nuclear magnetic resonance spectroscopy in vivo,” *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 34, no. 3, pp. 301 – 326, 1999.

[64] K. Wong, L. Markillie, and J. Saffer, “A novel method for producing partial restriction digestion of dna fragments by pcr with 5-methyl-ctp,” *Nucleic Acids Research*, vol. 25, no. 20, pp. 4169–4171, Oct. 1997.