# Refined Lower Bounds for Nearest Neighbor Condensation

Chitnis, Rajesh

# Refined Lower Bounds for Nearest Neighbor Condensation

**Rajesh Chitnis** RAJESHCHITNIS@GMAIL.COM
*School of Computer Science, University of Birmingham, UK*

**Editors:** Sanjoy Dasgupta and Nika Haghtalab

## Abstract

One of the most commonly used classification techniques is the nearest neighbor rule: given a training set $T$ of labeled points in a metric space $(\mathcal{X}, \rho)$, a new unlabeled point $x \in \mathcal{X}$ is assigned the label of its nearest neighbor in $T$. To improve both the space & time complexity of this classification, it is desirable to reduce the size of the training set without compromising too much on the accuracy of the classification. Hart (1968) formalized this as the NEAREST NEIGHBOR CONDENSATION (NNC) problem: find a subset $C \subseteq T$ of minimum size which is *consistent* with $T$, i.e., each point $t \in T$ has the same label as that of its nearest neighbor in $C$. This problem is known to be NP-hard (Wilfong, 1991), and the heuristics used in practice often have weak or no theoretical guarantees. We analyze this problem via the *refined* lens of parameterized complexity, and obtain strong lower bounds for the $k$-NNC-$(\mathbb{Z}^d, \ell_p)$ problem which asks if there is a consistent subset of size $\leq k$ for a given training set of size $n$ in the metric space $(\mathbb{Z}^d, \ell_p)$ for any $1 \leq p \leq \infty$:

- The $k$-NNC-$(\mathbb{Z}^d, \ell_p)$ problem is W[1]-hard parameterized by $k + d$, i.e., unless FPT = W[1], there is no $f(k, d) \cdot n^{O(1)}$ time algorithm for any computable function $f$.
- Under the Exponential Time Hypothesis (ETH), there is no $d \geq 2$ and computable function $f$ such that the $k$-NNC-$(\mathbb{Z}^d, \ell_p)$ problem can be solved in $f(k, d) \cdot n^{o(k^{1-1/d})}$ time.

The second lower bound shows that there is a so-called (Marx and Sidiropoulos, 2014) "limited blessing of low-dimensionality": for small $d$ some improvement *might be* possible over the brute-force $n^{O(k)}$ time algorithm, but as $d$ becomes large the brute-force algorithm becomes asymptotically optimal. It also shows that the is the $n^{O(\sqrt{k})}$ time algorithm of Biniaz et al. (2019) for $k$-NNC-$(\mathbb{R}^2, \ell_2)$ is asymptotically tight. Our lower bounds on the fine-grained complexity of NEAREST NEIGHBOR CONDENSATION in a sense justify the use of heuristics in practice, even though they have weak or no theoretical guarantees.

**Keywords:** nearest neighbor condensation, parameterized complexity, exponential time hypothesis

## 1. Introduction

**Motivation and Problem Statement:** The area of supervised learning tackles the problem of learning from a given labeled training set to correctly label new inputs. This can be done in two ways: either by classification or regression. One of the oldest and most commonly used classification techniques is the nearest neighbor rule (Fix and Hodges, 1951): given a training set $T$ of labeled points in a metric space $(\mathcal{X}, \rho)$, a new unlabeled point $x \in \mathcal{X}$ is assigned the label of its nearest neighbor in $T$. In addition to its simplicity, the nearest neighbor classification rule is also known to demonstrate good classification accuracy both in theory and practice (Cover and Hart, 1967; Devroye, 1981; Kontorovich and Weiss, 2015; Stone, 1977).

However, a naive implementation of the nearest neighbor classifier also has some disadvantages: in addition to having to store the entire training set in memory (Krauthgamer and Lee, 2004), there are some other issues such as high running time (Clarkson, 1994) and overfitting due to infinite VC dimension (Shalev-Shwartz and Ben-David, 2014). To improve both the space & time complexity

of nearest neighbor classification, it is highly desirable to reduce the size of the training set without compromising too much on the accuracy of the classification. This was formalized by Hart (1968) as the NEAREST NEIGHBOR CONDENSATION (NNC) problem: find a subset $C \subseteq T$ of minimum size which is *consistent* with the training set $T$, i.e., each point $t \in T$ has the same label as that of its nearest neighbor in $C$. The idea behind the NEAREST NEIGHBOR CONDENSATION problem is that given a large training set $T$ we can condense it into a smaller set $C \subseteq T$ which classifies the entire training set $T$ correctly: in this case, we could then use the smaller set $C$ for the classification instead of the original training set $T$. The NEAREST NEIGHBOR CONDENSATION problem is known to be NP-hard even for the case when there are only two labels (Khodamoradi et al., 2018; Wilfong, 1991; Zukhba, 2010).

**Related Work:** Most of the algorithmic research on the NEAREST NEIGHBOR CONDENSATION problem has focused on heuristics: indeed several heuristics have been designed such as CNN (Hart, 1968), FCNN (Angiulli, 2005), MCNN (Devi and Murty, 2002), RNN (Gates, 1972), SNN (Ritter et al., 1975), MSS (Barandela et al., 2005), RSS and VSS (Flores-Velazco and Mount, 2019), etc. We refer the interested reader to the surveys (Toussaint, 2002; Jankowski and Grochowski, 2004; Wilson and Martinez, 2000) for more details on these heuristics. These heuristics run in time which is quadratic or cubic in the size of the training set $T$, but despite significant experimental analysis (García et al., 2012) no guarantees were known on their performance. Only recently, there has been some work (Flores-Velazco and Mount, 2019; Flores-Velazco, 2020) which showed weak guarantees on the size of the consistent subset arising from these existing (and some new) heuristics. Gottlieb et al. (2014) designed a polynomial time algorithm called NET which computes an almost-tight[1] approximation of the minimum consistent subset. However, NET is often outperformed in practice by many of the aforementioned heuristics for the NEAREST NEIGHBOR CONDENSATION problem.

In addition to approximation algorithms, a popular algorithmic approach to cope with NP-hardness is via the refined lens of parameterized algorithms & complexity. Given that NP-hardness is only a worst-case intractability result, the paradigm of parameterized algorithms aims to analyze the effect of various relevant parameters of the problem (other than input size) on the running time. For the NEAREST NEIGHBOR CONDENSATION problem, a natural parameter is the size $k$ of the consistent subset we are looking for. This leads to the following problem:

$k$-**NNC-**$(\mathcal{X}, \rho)$
*Input*: A metric space $(\mathcal{X}, \rho)$, a set $T$ of $n$ points in $\mathcal{X}$ and a labeling function $\textsc{Label} : T \to [r]$ for some $r \in \mathbb{N}$
*Parameter*: An integer $k$ such that $k \geq r$
*Question*: Does there exists a subset $C \subseteq T$ of size $\leq k$ which is consistent with $T$, i.e., for every $t \in T$ if $x := \text{argmin}_{c \in C} \, \texttt{dist}_\rho(t, c)$ then $\textsc{Label}(x) = \textsc{Label}(t)$.

Note that the brute-force algorithm for the $k$-NNC-$(\mathcal{X}, \rho)$ problems runs in $n^{O(k)}$ time by enumerating all $\binom{n}{k}$ subsets of $k$ points, and then for each such choice checking[2] in $n^{O(1)}$ time if it forms a consistent subset. Algorithms which run in $f(k) \cdot |T|^{O(1)}$ time for some computable function $f$ are known as fixed-parameter tractable (FPT) algorithms. Two other well-known complexity classes

1. The algorithm of Gottlieb et al. (2014) works for any number of labels while their lower bound holds even there are only two labels.
2. This upper bound assumes that the distance between a pair of points can be computed *quickly*.

from parameterized complexity are W[1] and W[2]: these are the class of problems which can be reduced in FPT time to the INDEPENDENT SET and DOMINATING SET problems (respectively) where the parameter is the solution size. The standard hypothesis of parameterized complexity is that FPT $\subset$ W[1] $\subset$ W[2], i.e., each containment is strict. We refer the interested reader to Cygan et al. (2015) for more background about parameterized algorithms and complexity. Banerjee et al. (2018) showed lower bounds for the $k$-NNC-$(\mathcal{X}, \rho)$ problem when $(\mathcal{X}, \rho)$ is a graph metric: the problem is W[2]-hard, and moreover under the Exponential Time Hypothesis (ETH) (Impagliazzo and Paturi, 2001; Impagliazzo et al., 2001) has no $f(k) \cdot n^{o(k)}$ algorithm for any function $f$.

**Our Results:** In this paper, we study the parameterized complexity of the $k$-NNC-$(\mathcal{X}, \rho)$ problem in settings of practical relevance: a natural formulation used frequently in applications is to view the inputs as vectors in say $\mathbb{R}^d$ for some $d \geq 1$ and the distance function is given by the $\ell_p$-norm for some $1 \leq p \leq \infty$. The only prior result for NNC in this setting (that we are aware of) is the $n^{O(\sqrt{k})}$ time algorithm of Biniaz et al. (2019) for $k$-NNC-$(\mathbb{R}^2, \ell_2)$.

We obtain two lower bounds for the running time of the NEAREST NEIGHBOR CONDENSATION problem when the points are located in $\mathbb{Z}^d$ (this makes the lower bounds stronger) and the distances are measured using the $\ell_p$-norm for some $1 \leq p \leq \infty$:

**Theorem 1** For any $1 \leq p \leq \infty$, the $k$-NNC-$(\mathbb{Z}^d, \ell_p)$ problem is W[1]-hard parameterized by $k + d$, i.e., unless FPT = W[1], there is no $f(k, d) \cdot n^{O(1)}$ time algorithm for any computable function $f$.

We prove Theorem 1 by designing a parameterized reduction from the $d$-dimensional geometric $\geq$-CSP problem defined by Marx and Sidiropoulos (2014). Using the same reduction, we also obtain a stronger lower bound under a stronger (but still well-believed) hypothesis than FPT $\neq$ W[1]: this is the Exponential Time Hypothesis[3] (ETH) (Impagliazzo and Paturi, 2001; Impagliazzo et al., 2001).

**Theorem 2** For any $d \geq 2$, under the Exponential Time Hypothesis (ETH), the $k$-NNC-$(\mathbb{Z}^d, \ell_p)$ problem cannot be solved in $f(k) \cdot n^{o(k^{1-1/d})}$ time where $f$ is any computable function, $n$ is the total number of points in the training set, $k$ is the size of the consistent subset and $1 \leq p \leq \infty$.

Theorem 2 shows that there is a so-called *"limited blessing of low-dimensionality"* Marx and Sidiropoulos (2014): for small $d$ some improvement *might be* possible over the brute-force $n^{O(k)}$ time algorithm, but as $d$ becomes large the brute-force algorithm becomes asymptotically optimal. Note that Theorem 2 also implies that the $n^{O(\sqrt{k})}$ time algorithm of Biniaz et al. (2019) for $k$-NNC-$(\mathbb{R}^2, \ell_2)$ is asymptotically optimal.

Both of our lower bounds (Theorem 1 and Theorem 2) apply to instances of the NEAREST NEIGHBOR CONDENSATION problem which are of practical relevance, and hence justify the use of heuristics in practice even though the guarantees for them are weak or unknown. To keep the presentation simple, we present the proofs for $\ell_2$-metric: the small changes needed to extend the lower bounds to $\ell_p$-metrics for $1 \leq p \leq \infty$ are outlined in Remark 3.

3. ETH states that the 3-SAT problem cannot be solved in $2^{o(N)}$ time where $N$ is the number of variables. We refer to Lokshtanov et al. (2011) for more background on ETH and its consequences in parameterized complexity.

**Organization of the paper:** Section 2 starts by introducing the framework of Marx and Sidiropoulos Marx and Sidiropoulos (2014) which shows the hardness for solving the $d$-dimensional geometric $\geq$-CSP problem. Then in Section 2.1 we design a reduction from the $d$-dimensional geometric $\geq$-CSP problem to the $k$-NNC-$(\mathbb{Z}^d, \ell_p)$ problem. Correctness of the two directions of this reduction is shown in Section 2.2 and Section 2.3. Finally, we prove Theorem 1 and Theorem 2 in Section 2.4.

**Notation:** All vectors considered in this paper have length $d$. If $\mathbf{a}$ is a vector then for each $i \in [d]$ its $i^{\text{th}}$-coordinate is denoted by $\mathbf{a}[i]$. Addition and subtraction of vectors is denoted by $\oplus$ and $\ominus$ respectively. The $i^{\text{th}}$ unit vector is denoted by $\mathbf{e}_i$ and has $\mathbf{e}_i[i] = 1$ and $\mathbf{e}_i[j] = 0$ for each $j \neq i$. The set $\{1, 2, \ldots, n\}$ is denoted by $[n]$.

## 2. Lower Bounds for $k$-NNC-$(\mathbb{Z}^d, \ell_p)$

The goal of this section is to prove Theorem 1 and Theorem 2. We will do this via a reduction from the $d$-dimensional geometric $\geq$-CSP problem for which lower bounds were shown in Marx and Sidiropoulos (2014). We start with some necessary definitions before stating the formal theorems (Theorem 3 and Theorem 4) that will be used to obtain our lower bounds.

A Constraint Satisfaction Problem (CSPs) is an abstract formulation which captures several important problems. In this paper, we will work with a subclass of CSPs called binary CSPs which we define below:

**Definition 1** An instance of a binary constraint satisfaction problem (CSP) is a triple $\mathcal{I} = (V, D, C)$ where $V$ is a set of variables, $D$ is a domain of values and $C$ is a set of constraints. There are two types of constraints:

- *Unary constraints*: For some $v \in V$ there is a unary constraint $\langle v, R_v \rangle$ where $R_v \subseteq D$.
- *Binary constraints*: For some $u, v \in V$ there is a binary constraint $\langle (u, v), R_{u,v} \rangle$ where $R_{u,v} \subseteq D \times D$.

Given a CSP instance $\mathcal{I} = (V, D, C)$ the main question of interest is whether there exists a satisfying assignment for it, i.e., a function $f : V \to D$ such that all the constraints are satisfied. For a binary CSP, a satisfying assignment $f$ has the property that for each unary constraint $\langle v, R_v \rangle$ we have $f(v) \in R_v$ and for each binary constraint $\langle (u, v), R_{u,v} \rangle$ we have $\big(f(u), f(v)\big) \in R_{u,v}$.

The constraint graph of a CSP instance $\mathcal{I} = (V, D, C)$ is an undirected graph $G_{\mathcal{I}}$ whose vertex set is $V$ and the adjacency relation is defined as follows: two vertices $u, v \in V$ are adjacent in $G_{\mathcal{I}}$ if there is a constraint in $\mathcal{I}$ which contains both $u$ and $v$. The size $|\mathcal{I}|$ of a binary CSP $\mathcal{I} = (V, D, C)$ is the combined size of the variables, domain and the constraints. With appropriate preprocessing (for example, combining different constraints on the same variables) we can assume that $|\mathcal{I}| = \big(|V| + |D|\big)^{O(1)}$.

Marx and Sidiropoulos (2014) observed that considering binary CSPs whose primal graph is a subgraph of the $d$-dimensional grid is useful in showing lower bounds for geometric problems in $d$-dimensions.

**Definition 2** The $d$-dimensional grid $\mathrm{R}[\kappa, d]$ is an undirected graph with vertex set $[\kappa]^d$ and the adjacency relation is as follows: two vertices $\mathbf{a}, \mathbf{b} \in [\kappa]^d$ have an edge between them if and only if $\sum_{i=1}^{d} \big|\mathbf{a}[i] - \mathbf{b}[i]\big| = 1$.

**Definition 3** A $d$-dimensional geometric $\geq$-CSP $\mathcal{I} = (V, D, C)$ is a binary CSP whose

- set of variables $V$ is a subset of $\mathrm{R}[\kappa, d]$ for some $\kappa \geq 1$
- domain is $[N]^d$ for some integer $N \geq 1$
- constraint graph $G_{\mathcal{I}}$ is an induced subgraph of $\mathrm{R}[\kappa, d]$
- binary constraints are of the following type: if $\mathbf{a}, \mathbf{a}' \in V$ such that $\mathbf{a}' = \mathbf{a} \oplus \mathbf{e}_i$ for some $i \in [d]$ then there is a binary constraint $\langle (\mathbf{a}, \mathbf{a}'), R_{\mathbf{a},\mathbf{a}'} \rangle$ where we have $R_{\mathbf{a},\mathbf{a}'} = \left\{ (\mathbf{x}, \mathbf{y}) \in \mathrm{R}_{\mathbf{a}} \times \mathrm{R}_{\mathbf{a}'} \mid \mathbf{x}[i] \geq \mathbf{y}[i] \right\}$

**Remark 1** The problem defined by Marx and Sidiropoulos (2014) is actually $d$-dimensional geometric $\leq$-CSP which has $\leq$-constraints instead of the $\geq$-constraints. However, for each $\mathbf{a} \in V$ by replacing each unary constraint $\mathbf{x} \in R_{\mathbf{a}}$ by $\mathbf{y}$ such that $\mathbf{y}[i] = N + 1 - \mathbf{x}[i]$ for each $i \in [d]$, it is easy to see that $d$-dimensional geometric $\leq$-CSP and $d$-dimensional geometric $\geq$-CSP are equivalent.

We now state the two lower bounds for the $d$-dimensional geometric $\geq$-CSP problem which were shown by Marx and Sidiropoulos (2014).

**Theorem 3** (Implicit in Marx and Sidiropoulos (2014); Cygan et al. (2015)) The $d$-dimensional geometric $\geq$-CSP problem is W[1]-hard parameterized by $|V| + d$.

**Proof** We reduce from the $(\kappa \times \kappa)$-GRID-TILING-$\geq$ problem defined by Marx and Sidiropoulos (2014) :

> $(\kappa \times \kappa)$-**GRID-TILING-$\geq$**
> *Input*: integers $\kappa, N$, and a collection $\mathcal{S}$ of $\kappa^2$ non-empty sets $S[x, y] \subseteq [N] \times [N]$ where $1 \leq x, y \leq \kappa$.
> *Question*: for each $1 \leq x, y \leq \kappa$ does there exist a value $\gamma_{x,y} \in S[x, y]$ such that
> - if $\gamma_{x,y} = (a, b)$ and $\gamma_{x+1,y} = (a', b')$ then $a \geq a'$
> - if $\gamma_{x,y} = (a, b)$ and $\gamma_{x,y+1} = (a', b')$ then $b \geq b'$

Observe that $(\kappa \times \kappa)$-GRID-TILING-$\geq$ is a special case of $d$-dimensional geometric $\geq$-CSP when $d = 2$ and $V = R[\kappa, 2]$: the unary constraints are given by the sets in $\mathcal{S}$, and the binary constraints are exactly what is the condition of $(\kappa \times \kappa)$-GRID-TILING-$\geq$. It is known[4] (Cygan et al., 2015, Theorem 14.30) that $(\kappa \times \kappa)$-GRID-TILING-$\geq$ is W[1]-hard parameterized by $k$. Combining this with the trivial reduction[5] from $d$-dimensional geometric $\geq$-CSP to $(d+1)$-dimensional geometric $\geq$-CSP, it follows that the $d$-dimensional geometric $\geq$-CSP is W[1]-hard parameterized by $|V| + d$. ■

The next result gives a lower bound on the running time (under ETH) for any algorithm that solves a $d$-dimensional geometric $\geq$-CSP:

**Theorem 4** (Marx and Sidiropoulos, 2014, Theorem 2.1) If for some fixed $d \geq 2$, there is an $f(|V|) \cdot |\mathcal{I}|^{o(|V|^{1-1/d})}$ time algorithm for solving a $d$-dimensional geometric $\geq$-CSP $\mathcal{I}$ for some computable function $f$, then the Exponential Time Hypothesis (ETH) fails.

4. The problem defined in Cygan et al. (2015) is actually $(\kappa \times \kappa)$-GRID-TILING-$\leq$ which has $\leq$-constraints instead of the $\geq$-constraints in the $(\kappa \times \kappa)$-GRID-TILING-$\geq$ problem. However, by replacing each $(a, b) \in S[x, y]$ by $(N + 1 - a, N + 1 - b)$ for each $(x, y) \in [k] \times [k]$ it is easy to see that $(\kappa \times \kappa)$-GRID-TILING-$\leq$ and $(\kappa \times \kappa)$-GRID-TILING-$\geq$ are equivalent.
5. Set the $(d+1)^{th}$ co-ordinate of all variables to be 0

By showing the correctness of our reduction from $d$-dimensional geometric $\geq$-CSP to $k$-NNC-$(\mathbb{Z}^d, \ell_p)$, we will then be able to leverage the lower bounds from Theorem 3 and Theorem 4 to prove Theorem 1 and Theorem 2 respectively.

### 2.1. Construction of the $k$-NNC-$(\mathbb{R}^d, \ell_2)$ instance

Let $\mathcal{I} = (V, D, C)$ be an instance of $d$-dimensional geometric $\geq$-CSP. From this we will now construct an instance $\mathcal{U} = (\mathcal{P}, \text{LABEL})$ of $k$-NNC-$(\mathbb{R}^d, \ell_2)$ such that $\mathcal{I}$ has a satisfying assignment if and only if $\mathcal{U}$ has a consistent subset of size $|V|$.

Fix the following three values

$$\epsilon = \frac{1}{4}; \quad D^* = 2d \cdot N^2; \quad C^* = 2D^* + (N-1) \tag{1}$$

$$\text{Assume } d, N \geq 2 \text{ and so Eqn. 1} \Rightarrow D^* \geq 4N^2 \geq 16 \tag{2}$$

#### 2.1.1. Adding internal points:

For each $\mathbf{a} = (a_1, a_2, \ldots, a_d) \in V$ we define a set of points denoted by $\text{INTERNAL}_1(\mathbf{a})$ as follows:

- The $\texttt{temp-origin}(\mathbf{a})$ is $1^d = (1, 1, \ldots, 1)$.
- For each $\mathbf{x} \in R_{\mathbf{a}} \subseteq [N]^d$ we add a point located at $\mathbf{x} = \texttt{temp-origin}(\mathbf{a}) \oplus \left(\mathbf{x} \ominus 1^d\right)$.

We now perform the following three operations (in that order) to obtain our final set of points $\mathcal{P}$:

#### 2.1.2. Mirroring:

For each $i \in [\kappa]$ we define the function $\texttt{flip}_i : [N] \to [N]$ as follows: for each $q \in [N]$

$$\texttt{flip}_i(q) = \begin{cases} N + 1 - q & \text{if } i \text{ is even} \\ q & \text{if } i \text{ is odd} \end{cases} \tag{3}$$

**Observation 1** Note that for each $i \in [\kappa]$ and each $q \in [N]$ we have $\texttt{flip}_i\big(\texttt{flip}_i(q)\big) = q$.

For each $\mathbf{a} \in V$, we make "mirroring" changes to all the points of $\text{INTERNAL}_1(\mathbf{a})$ as follows:

- If $\mathbf{x} \in \text{INTERNAL}_1(\mathbf{a})$ then we replace it with $\mathbf{y}$ where $\mathbf{y}[i] = \texttt{flip}_{\mathbf{a}[i]}\big(\mathbf{x}[i]\big)$ for each $i \in [d]$.

We call this set of points as $\text{INTERNAL}_2(\mathbf{a})$.

#### 2.1.3. Translation:

We now fix the location of the origin of each grids by translation as follows: for each $\mathbf{a} \in V$ set

$$\texttt{origin}(\mathbf{a}) = \texttt{temp-origin}(\mathbf{a}) \oplus C^* \cdot \left(\mathbf{a} \ominus 1^d\right) = 1^d \oplus C^* \cdot \left(\mathbf{a} \ominus 1^d\right) \tag{4}$$

Note that this also shifts all points of $\text{INTERNAL}_2(\mathbf{a})$ accordingly: each point $\mathbf{y} \in \text{INTERNAL}_2(\mathbf{a})$ is shifted to the point $\texttt{origin}(\mathbf{a}) \oplus \left(\mathbf{y} \ominus 1^d\right)$. We denote this new set of points by $\text{INTERNAL}_3(\mathbf{a})$.

### 2.1.4. Adding border vertices:

For each $\mathbf{a} \in V$, we define a set of "border" points by adding points corresponding to the adjacencies in $G_{\mathcal{I}}$. Since $G_{\mathcal{I}}$ is an induced subgraph of the $d$-dimensional grid $\mathrm{R}[\kappa, d]$, every edge in $G_{\mathcal{I}}$ is of the following form: there exist $\mathbf{b} \in V$ and $j \in [d]$ such that the endpoints of the edge are $\mathbf{b}$ and $\mathbf{b} \oplus \mathbf{e}_j$.

We have two cases depending on the parity of $\mathbf{a}[i]$:

1. $\mathbf{a}[i]$ is odd:
   If $\mathbf{a}$ and $(\mathbf{a} \oplus \mathbf{e}_i)$ form an edge in $G_{\mathcal{I}}$ then add the following point to the set $\textsc{Border}_{+i}(\mathbf{a})$
   - $\mathtt{mid}_{\mathbf{a}}^{+i} := \mathtt{origin}(\mathbf{a}) \oplus \mathbf{e}_i \cdot \big((N-1) + (D^* - \epsilon)\big)$

   If $\mathbf{a}$ and $(\mathbf{a} \ominus \mathbf{e}_i)$ form an edge in $G_{\mathcal{I}}$ then add the following point to the set $\textsc{Border}_{-i}(\mathbf{a})$
   - $\mathtt{mid}_{\mathbf{a}}^{-i} := \mathtt{origin}(\mathbf{a}) \ominus \mathbf{e}_i \cdot \big((D^* - \epsilon)\big)$
2. $\mathbf{a}[i]$ is even:
   If $\mathbf{a}$ and $(\mathbf{a} \oplus \mathbf{e}_i)$ form an edge in $G_{\mathcal{I}}$ then add the following two points to the set $\textsc{Border}_{+i}(\mathbf{a})$
   - $\mathtt{plus}_{\mathbf{a}}^{+i} := \mathtt{origin}(\mathbf{a}) \oplus \mathbf{e}_i \cdot \big((N-1) + (D^* + \epsilon - N)\big) \oplus (1^d \ominus \mathbf{e}_i) \cdot N$
   - $\mathtt{minus}_{\mathbf{a}}^{+i} := \mathtt{origin}(\mathbf{a}) \oplus \mathbf{e}_i \cdot \big((N-1) + (D^* + \epsilon - N)\big) \ominus (1^d \ominus \mathbf{e}_i) \cdot N$

   If $\mathbf{a}$ and $(\mathbf{a} \ominus \mathbf{e}_i)$ form an edge in $G_{\mathcal{I}}$ then add the following two points to the set $\textsc{Border}_{-i}(\mathbf{a})$
   - $\mathtt{plus}_{\mathbf{a}}^{-i} := \mathtt{origin}(\mathbf{a}) \ominus \mathbf{e}_i \cdot (D^* + \epsilon - N) \oplus (1^d \ominus \mathbf{e}_i) \cdot N$
   - $\mathtt{minus}_{\mathbf{a}}^{-i} := \mathtt{origin}(\mathbf{a}) \ominus \mathbf{e}_i \cdot (D^* + \epsilon - N) \ominus (1^d \ominus \mathbf{e}_i) \cdot N$

   For each $\mathbf{a} \in V$, we define

$$\textsc{Border}(\mathbf{a}) = \bigcup_{i=1}^{d} \Big( \textsc{Border}_{+i}(\mathbf{a}) \cup \textsc{Border}_{-i}(\mathbf{a}) \Big) \tag{5}$$

Hence, it follows that $\big|\textsc{Border}(\mathbf{a})\big| \leq 2 \cdot (2d) = 4d$ for each $\mathbf{a} \in V$. The final set of points is given by

$$\mathcal{P} := \bigcup_{\mathbf{a} \in V} \mathcal{P}(\mathbf{a}) \text{where } \mathcal{P}(\mathbf{a}) := \textsc{Internal}_3(\mathbf{a}) \bigcup \textsc{Border}(\mathbf{a}) \tag{6}$$

The set of points given by $\bigcup_{\mathbf{a} \in V} \textsc{Internal}_3(\mathbf{a})$ is called as 'internal" vertices, and set of points given by $\bigcup_{\mathbf{a} \in V} \textsc{Border}(\mathbf{a})$ is called as 'border" vertices. Note that we add one point corresponding to unary constraint on each variable and at most 3 points corresponding to each edge in $G_{\mathcal{I}}$. Hence, the total number of points $n$ in the instance $\mathcal{U}$ is $\leq |C| + 3 \cdot |V|^2 = |\mathcal{I}|^{O(1)}$ where $|\mathcal{I}| = |V| + |D| + |C|$.

### 2.1.5. Labeling the points:

Fix any bijective function $\mathtt{Bij} : V \to \big[|V|\big]$. We now define a labeling/coloring function $\textsc{Label} : \mathcal{P} \to \big[|V|\big]$ as follows: for each $\mathbf{a} \in V$, all points from the set $\mathcal{P}(\mathbf{a})$ are given the label $\mathtt{Bij}(\mathbf{a})$. Hence, we have a total of $|V|$ labels given by the set $\big[|V|\big]$.

This completes the construction of the instance $(\mathcal{P}, \textsc{Label})$ of $k$-NNC-$(\mathbb{R}^d, \ell_2)$.

**Remark 2** Since $\epsilon = 1/4$ (Eqn. 1), it follows that multiplying each coordinate by four ensures that all points in $\mathcal{P}$ actually lie in $\mathbb{Z}^d$. This scaling uniformly increases all pairwise distances by the same factor, and hence does not affect the correctness of the reduction.

### 2.2. $\mathcal{U}$ has a consistent subset of size $|V| \Rightarrow \mathcal{I}$ has a satisfying assignment

Let $\mathcal{P}' \subseteq \mathcal{P}$ be a consistent subset of size $|V|$ for the instance $\mathcal{U} = (\mathcal{P}, \text{LABEL})$ of $k$-NNC-$(\mathbb{R}^d, \ell_2)$. We will now show that the instance $\mathcal{I} = (V, D, C)$ of $d$-dimensional geometric $\geq$-CSP has a satisfying assignment.

Since there are $|V|$ different labels, we are forced to pick exactly one point from each color into $\mathcal{P}'$, i.e.,

$$\left|\mathcal{P}' \cap \mathcal{P}(\mathbf{a})\right| = 1 \quad \text{for each} \quad \mathbf{a} \in V \tag{7}$$

First we start with a preliminary lemma:

**Lemma 1** Let $\mathbf{a} \in V$ and $i \in [d]$.
(1) For any point $p \in \mathcal{P}(\mathbf{a}) \setminus \text{BORDER}_{+i}(\mathbf{a})$, the absolute value of the difference in the $i^{th}$-coordinates of $p$ and any point in $\text{BORDER}_{+i}(\mathbf{a})$ is at least $(D^* - \epsilon) - 2N$
(2) For any point $p \in \mathcal{P}(\mathbf{a}) \setminus \text{BORDER}_{-i}(\mathbf{a})$, the absolute value of the difference in the $i^{th}$-coordinates of $p$ and any point in $\text{BORDER}_{-i}(\mathbf{a})$ is at least $(D^* - \epsilon) - 2N$

**Proof** Let $\lambda$ be the $i^{th}$-coordinate of $\texttt{origin}(\mathbf{a})$. We start with proving the first part of the lemma:

**Claim 1** The $i^{th}$-coordinate of any point from $\text{BORDER}_{+i}(\mathbf{a})$ is at least $\big(\lambda + (D^* + \epsilon) - 1\big)$

**Proof** If $a_i$ is odd, then the $i^{th}$-coordinate of the only point in $\text{BORDER}_{+i}(\mathbf{a})$ (viz. $\texttt{mid}_{\mathbf{a}}^{+i}$) is equal to $\lambda + (N-1) + (D^* - \epsilon)$ which is greater than $\big(\lambda + (D^* + \epsilon) - 1\big)$ since $N \geq 2$ (Eqn. 2) and $\epsilon = 1/4$ (Eqn. 1). If $a_i$ is even, then the $i^{th}$-coordinate of both points in $\text{BORDER}_{+i}(\mathbf{a})$, viz. $\texttt{plus}_{\mathbf{a}}^{+i}$ and $\texttt{minus}_{\mathbf{a}}^{+i}$, is equal to $\big(\lambda + (D^* + \epsilon) - 1\big)$. ■

Since $p \in \mathcal{P}(\mathbf{a}) \setminus \text{BORDER}_{+i}(\mathbf{a})$, by Eqn. 5 and Eqn. 6 it is sufficient to consider the following four cases:

- $p \in \text{INTERNAL}_3(\mathbf{a})$: By Section 2.1.1, the value of the $i^{th}$-coordinate of $p$ is at most $\lambda + (N-1)$. By Claim 1, the absolute value of the difference in the $i^{th}$-coordinates of $p$ and any point in $\text{BORDER}_{+i}(\mathbf{a})$ is at least $D^* + \epsilon - N$ which is greater than $(D^* - \epsilon) - 2N$ since $N \geq 2$ (Eqn. 2) and $\epsilon = 1/4$ (Eqn. 1).
- $p \in \text{BORDER}_{-i}(\mathbf{a})$: By Section 2.1.4, the value of the $i^{th}$-coordinate of $p$ is $< \lambda$. By Claim 1, the absolute value of the difference in the $i^{th}$-coordinates of $p$ and any point in $\text{BORDER}_{+i}(\mathbf{a})$ is at least $D^* + \epsilon - 1$ which is greater than $(D^* - \epsilon) - 2N$ since $N \geq 2$ (Eqn. 2) and $\epsilon = 1/4$ (Eqn. 1).
- $p \in \text{BORDER}_{+j}(\mathbf{a})$ for some $j \in [d]$ such that $j \neq i$: If $a_j$ is odd then the only point in $\text{BORDER}_{+j}(\mathbf{a})$ is $\texttt{mid}_{\mathbf{a}}^{+j}$ whose $i^{th}$-coordinate is $\lambda$. By Claim 1, the absolute value of the difference in the $i^{th}$-coordinates of $p$ and any point in $\text{BORDER}_{+i}(\mathbf{a})$ is at least $D^* + \epsilon - 1$ which is greater than $(D^* - \epsilon) - 2N$ since $N \geq 2$ (Eqn. 2) and $\epsilon = 1/4$ (Eqn. 1).
  If $a_j$ is even then the only points in $\text{BORDER}_{+j}(\mathbf{a})$ are $\texttt{plus}_{\mathbf{a}}^{+j}$ and $\texttt{minus}_{\mathbf{a}}^{+j}$ whose $i^{th}$-coordinates are $\lambda + N$ and $\lambda - N$ respectively. By Claim 1, the absolute value of the difference in the $i^{th}$-coordinates of $p$ and any point in $\text{BORDER}_{+i}(\mathbf{a})$ is at least $\Big(\big(\lambda + (D^* + \epsilon) - 1\big) - \big(\lambda + N\big) = (D^* + \epsilon) - 1 - N$ which is greater than $(D^* - \epsilon) - 2N$ since $N \geq 2$ (Eqn. 2) and $\epsilon = 1/4$ (Eqn. 1).
- $p \in \text{BORDER}_{-j}(\mathbf{a})$ for some $j \in [d]$ such that $j \neq i$: The arguments for this case are exactly the same as the previous case.

We now prove the second part of the lemma:

**Claim 2** The $i^{th}$-coordinate of any point from $\textsc{Border}_{-i}(\mathbf{a})$ is at most $\big(\lambda - (D^* + \epsilon - N)\big)$

**Proof** If $a_i$ is odd, then the $i^{th}$-coordinate of the only point in $\textsc{Border}_{-i}(\mathbf{a})$ (viz. $\texttt{mid}_{\mathbf{a}}^{-i}$) is equal to $\lambda - (D^* - \epsilon)$ which is less than $\big(\lambda - (D^* + \epsilon - N)\big)$ since $N \geq 2$ (Eqn. 2) and $\epsilon = 1/4$ (Eqn. 1). If $a_i$ is even, then the $i^{th}$-coordinate of both points in $\textsc{Border}_{-i}(\mathbf{a})$, viz. $\texttt{plus}_{\mathbf{a}}^{-i}$ and $\texttt{minus}_{\mathbf{a}}^{-i}$, is equal to $\big(\lambda - (D^* + \epsilon - N)\big)$. ■

Since $p \in \mathcal{P}(\mathbf{a}) \setminus \textsc{Border}_{+i}(\mathbf{a})$, by Eqn. 5 and Eqn. 6 it is sufficient to consider the following four cases:

- $p \in \textsc{Internal}_3(\mathbf{a})$: By Section 2.1.1, the value of the $i^{th}$-coordinate of $p$ is at least $\lambda$. By Claim 2, the absolute value of the difference in the $i^{th}$-coordinates of $p$ and any point in $\textsc{Border}_{-i}(\mathbf{a})$ is at least $D^* + \epsilon - N$ which is greater than $(D^* - \epsilon) - 2N$ since $N \geq 2$ (Eqn. 2) and $\epsilon = 1/4$ (Eqn. 1).
- $p \in \textsc{Border}_{+i}(\mathbf{a})$: By Section 2.1.4, the value of the $i^{th}$-coordinate of $p$ is $> \lambda$. By Claim 2, the absolute value of the difference in the $i^{th}$-coordinates of $p$ and any point in $\textsc{Border}_{-i}(\mathbf{a})$ is at least $D^* + \epsilon - N$ which is greater than $(D^* - \epsilon) - 2N$ since $N \geq 2$ (Eqn. 2) and $\epsilon = 1/4$ (Eqn. 1).
- $p \in \textsc{Border}_{+j}(\mathbf{a})$ for some $j \in [d]$ such that $j \neq i$: If $a_j$ is odd then the only point in $\textsc{Border}_{+j}(\mathbf{a})$ is $\texttt{mid}_{\mathbf{a}}^{+j}$ whose $i^{th}$-coordinate is $\lambda$. By Claim 2, the absolute value of the difference in the $i^{th}$-coordinates of $p$ and any point in $\textsc{Border}_{-i}(\mathbf{a})$ is at least $D^* + \epsilon - N$ which is greater than $(D^* - \epsilon) - 2N$ since $N \geq 2$ (Eqn. 2) and $\epsilon = 1/4$ (Eqn. 1).
  If $a_j$ is even then the only points in $\textsc{Border}_{+j}(\mathbf{a})$ are $\texttt{plus}_{\mathbf{a}}^{+j}$ and $\texttt{minus}_{\mathbf{a}}^{+j}$ whose $i^{th}$-coordinates are $\lambda + N$ and $\lambda - N$ respectively. By Claim 1, the absolute value of the difference in the $i^{th}$-coordinates of $p$ and any point in $\textsc{Border}_{+i}(\mathbf{a})$ is at least $\big(\lambda - N\big) - \big(\lambda - (D^* + \epsilon - N)\big) = (D^* + \epsilon) - 2N$ which is greater than $(D^* - \epsilon) - 2N$ since $\epsilon = 1/4$ (Eqn. 1).
- $p \in \textsc{Border}_{-j}(\mathbf{a})$ for some $j \in [d]$ such that $j \neq i$: The arguments for this case are exactly the same as the previous case.

This concludes the proof of Lemma 1. ■

Next we show that $\mathcal{P}'$ cannot contain any border vertices.

**Lemma 2** $\mathcal{P}'$ contains no border vertices, i.e., $\mathcal{P}' \cap \textsc{Border}(\mathbf{a}) = \emptyset$ for each $\mathbf{a} \in V$.

**Proof** Suppose there exists some $\mathbf{a} \in V$ such that $\mathcal{P}' \cap \textsc{Border}(\mathbf{a}) \neq \emptyset$. Let $\mathbf{z} \in \big(\mathcal{P}' \cap \textsc{Border}(\mathbf{a})\big)$. From Eqn. 7 it follows that $\mathcal{P}' \cap \textsc{Border}(\mathbf{a}) = \{\mathbf{z}\}$. By Eqn. 5, there exists $i \in [d]$ such that either $\mathbf{z} \in \textsc{Border}_{+i}(\mathbf{a})$ or $\mathbf{z} \in \textsc{Border}_{-i}(\mathbf{a})$. Without loss of generality let $\mathbf{z} \in \textsc{Border}_{+i}(\mathbf{a})$: the argument for the other case is analogous. We need to consider the following two cases depending on the parity of $a_i$:

(1) $\mathbf{a}[i]$ is odd: By Section 2.1.4 and Eqn. 4, it follows that
  - $\mathbf{z} = \texttt{mid}_{\mathbf{a}}^{+i} = \texttt{origin}(\mathbf{a}) \oplus \mathbf{e}_i \cdot \big((N-1) + (D^* - \epsilon)\big)$
  - $\texttt{plus}_{\mathbf{a} \oplus \mathbf{e}_i}^{-i} = \texttt{origin}(\mathbf{a} \oplus \mathbf{e}_i) \ominus \mathbf{e}_i \cdot (D^* + \epsilon - N) \oplus (1^d \ominus \mathbf{e}_i) \cdot N = \texttt{mid}_{\mathbf{a}}^{+i} \oplus \mathbf{e}_i \cdot N \oplus (1^d \ominus \mathbf{e}_i) \cdot N$
  - $\texttt{minus}_{\mathbf{a} \oplus \mathbf{e}_i}^{-i} = \texttt{origin}(\mathbf{a} \oplus \mathbf{e}_i) \ominus \mathbf{e}_i \cdot (D^* + \epsilon - N) \ominus (1^d \ominus \mathbf{e}_i) \cdot N = \texttt{mid}_{\mathbf{a}}^{+i} \oplus \mathbf{e}_i \cdot N \ominus (1^d \ominus \mathbf{e}_i) \cdot N$

We now show that $|\mathcal{P}' \cap \mathcal{P}(\mathbf{a} \oplus \mathbf{e}_i)| \geq 2$ which is a contradiction to Eqn. 7.

**Claim 3** For any $p \in \mathcal{P}(\mathbf{a} \oplus \mathbf{e}_i)$ such that $p \neq \texttt{plus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}$ we have $\texttt{dist}\left(\texttt{mid}^{+i}_{\mathbf{a}}, \texttt{plus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}\right) < \texttt{dist}\left(p, \texttt{plus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}\right)$

**Proof** The proof follows from the following two observations:

– The distance between $\texttt{mid}^{+i}_{\mathbf{a}}$ and $\texttt{plus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}$ is less than that between $\texttt{plus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}$ and $\texttt{minus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}$ because

$$\begin{aligned}&\texttt{dist}\left(\texttt{mid}^{+i}_{\mathbf{a}}, \texttt{plus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}\right) = \sqrt{dN^2} \\ &< \sqrt{(d-1) \cdot (2N)^2} = \texttt{dist}\left(\texttt{plus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}, \texttt{minus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}\right)\end{aligned}$$

where we have used the fact that $d, N \geq 2$ (Eqn. 2),

– For any $q \in \mathcal{P}(\mathbf{a} + \mathbf{e}_i) \setminus \{\texttt{plus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}, \texttt{minus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}\}$ we have

$$\begin{aligned}&\texttt{dist}\left(q, \texttt{plus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}\right) \\ &\geq (D^* - \epsilon) - 2N && \text{(by Lemma 1)} \\ &> \sqrt{dN^2} && \text{(from Eqn. 1 and Eqn. 2)} \\ &= \texttt{dist}\left(\texttt{mid}^{+i}_{\mathbf{a}}, \texttt{plus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}\right)\end{aligned}$$

■

Since $\texttt{mid}^{+i}_{\mathbf{a}} \in \mathcal{P}'$ and $\textsc{Label}\left(\texttt{mid}^{+i}_{\mathbf{a}}\right) = \texttt{Bij}(\mathbf{a}) \neq \texttt{Bij}(\mathbf{a} + \mathbf{e}_i) = \textsc{Label}\left(\texttt{plus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}\right)$, it follows from Claim 3 that $\texttt{plus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i} \in \mathcal{P}'$.

**Claim 4** For any $p \in \mathcal{P}(\mathbf{a} \oplus \mathbf{e}_i)$ such that $p \neq \texttt{minus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}$ we have $\texttt{dist}\left(\texttt{mid}^{+i}_{\mathbf{a}}, \texttt{minus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}\right) < \texttt{dist}\left(p, \texttt{minus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}\right)$

**Proof** The proof of this claim is exactly the same as that of Claim 3. ■

Since $\texttt{mid}^{+i}_{\mathbf{a}} \in \mathcal{P}'$ and $\textsc{Label}\left(\texttt{mid}^{+i}_{\mathbf{a}}\right) = \texttt{Bij}(\mathbf{a}) \neq \texttt{Bij}(\mathbf{a} + \mathbf{e}_i) = \textsc{Label}\left(\texttt{minus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}\right)$, it follows from Claim 4 that $\texttt{minus}^{i}_{\mathbf{a} \oplus \mathbf{e}_i} \in \mathcal{P}'$. Therefore, we have $\left\{\texttt{plus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}, \texttt{minus}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}\right\} \subseteq \left(\mathcal{P}' \cap \mathcal{P}(\mathbf{a} \oplus \mathbf{e}_i)\right)$ which contradicts Eqn. 7.

(2) $\overline{\mathbf{a}[i] \text{ is even}}$: By Section 2.1.4 and Eqn. 4, it follows that

– $\texttt{mid}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i} = \texttt{origin}(\mathbf{a} \oplus \mathbf{e}_i) \ominus \mathbf{e}_i \cdot \left((D^* - \epsilon)\right)$

– $\texttt{plus}^{+i}_{\mathbf{a}} = \texttt{origin}(\mathbf{a}) \oplus \mathbf{e}_i \cdot \left((N-1) + (D^* + \epsilon - N)\right) \oplus (1^d \ominus \mathbf{e}_i) \cdot N = \texttt{mid}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i} \ominus \mathbf{e}_i \cdot N \oplus (1^d \ominus \mathbf{e}_i) \cdot N$

– $\texttt{minus}^{+i}_{\mathbf{a}} = \texttt{origin}(\mathbf{a}) \oplus \mathbf{e}_i \cdot \left((N-1) + (D^* + \epsilon - N)\right) \ominus (1^d \ominus \mathbf{e}_i) \cdot N = \texttt{mid}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i} \ominus \mathbf{e}_i \cdot N \ominus (1^d \ominus \mathbf{e}_i) \cdot N$

By Section 2.1.4, we have either $\mathbf{z} = \texttt{plus}^{+i}_{\mathbf{a}}$ or $\mathbf{z} = \texttt{minus}^{+i}_{\mathbf{a}}$. Let $\mathbf{z} = \texttt{plus}^{+i}_{\mathbf{a}}$: the case when $\mathbf{z} = \texttt{minus}^{+i}_{\mathbf{a}}$ is analogous. We will now show that $|\mathcal{P}' \cap \mathcal{P}(\mathbf{a})| \geq 2$ which is a contradiction to Eqn. 7.

**Claim 5** For any $p \in \mathcal{P}(\mathbf{a} \oplus \mathbf{e}_i)$ such that $p \neq \texttt{mid}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}$ we have $\texttt{dist}\left(\texttt{mid}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}, \texttt{plus}^{+i}_{\mathbf{a}}\right) < \texttt{dist}\left(p, \texttt{mid}^{-i}_{\mathbf{a} \oplus \mathbf{e}_i}\right)$

**Proof** Let $p \in \mathcal{P}(\mathbf{a} \oplus \mathbf{e}_i) \setminus \{\texttt{mid}^{-i}_{\mathbf{a}\oplus\mathbf{e}_i}\}$. Since $\mathbf{a}[i]$ is even, it follows that $\textsc{Border}_{-i}(\mathbf{a}\oplus\mathbf{e}_i) = \{\texttt{mid}_{\mathbf{a}\oplus\mathbf{e}_i}(-i)\}$. Hence, we have

$$\begin{aligned}
&\texttt{dist}(p, \texttt{mid}^{-i}_{\mathbf{a}\oplus\mathbf{e}_i}) \\
&\geq (D^* - \epsilon) - 2N && \text{(by Lemma 1)} \\
&> \sqrt{dN^2} && \text{(from Eqn. 1 and Eqn. 2)} \\
&= \texttt{dist}(\texttt{mid}^{-i}_{\mathbf{a}\oplus\mathbf{e}_i}, \texttt{plus}^{+i}_{\mathbf{a}})
\end{aligned}$$

■

Since $\texttt{plus}^{+i}_{\mathbf{a}} \in \mathcal{P}'$ and $\textsc{Label}\big(\texttt{mid}^{-i}_{\mathbf{a}\oplus\mathbf{e}_i}\big) = \texttt{Bij}(\mathbf{a} \oplus \mathbf{e}_i) \neq \texttt{Bij}(\mathbf{a}) = \textsc{Label}\big(\texttt{plus}^{+i}_{\mathbf{a}}\big)$, it follows from Claim 5 that $\texttt{mid}^{-i}_{\mathbf{a}\oplus\mathbf{e}_i} \in \mathcal{P}'$.

**Claim 6** For any $p \in \mathcal{P}(\mathbf{a})$ such that $p \neq \texttt{minus}^{+i}_{\mathbf{a}}$ we have $\texttt{dist}\big(\texttt{mid}^{-i}_{\mathbf{a}\oplus\mathbf{e}_i}, \texttt{minus}^{+i}_{\mathbf{a}}\big) < \texttt{dist}\big(p, \texttt{minus}^{+i}_{\mathbf{a}}\big)$

**Proof** The proof follows from the following two observations:

– The distance between $\texttt{mid}^{-i}_{\mathbf{a}\oplus\mathbf{e}_i}$ and $\texttt{minus}^{+i}_{\mathbf{a}}$ is less than that between $\texttt{plus}^{-i}_{\mathbf{a}\oplus\mathbf{e}_i}$ and $\texttt{minus}^{+i}_{\mathbf{a}}$ because

$$\begin{aligned}
&\texttt{dist}\big(\texttt{mid}^{-i}_{\mathbf{a}\oplus\mathbf{e}_i}, \texttt{minus}^{+i}_{\mathbf{a}}\big) = \sqrt{dN^2} \\
&< \sqrt{(d-1)\cdot(2N)^2} && \text{(from Eqn. 2)} \\
&= \texttt{dist}\big(\texttt{plus}^{-i}_{\mathbf{a}\oplus\mathbf{e}_i}, \texttt{minus}^{-i}_{\mathbf{a}\oplus\mathbf{e}_i}\big)
\end{aligned}$$

– $\forall\, q \in \mathcal{P}(\mathbf{a}) \setminus \{\texttt{plus}^{+i}_{\mathbf{a}}, \texttt{minus}^{+i}_{\mathbf{a}}\}$, we have

$$\begin{aligned}
&\texttt{dist}\big(q, \texttt{minus}^{+i}_{\mathbf{a}}\big) \\
&\geq (D^* - \epsilon) - 2N && \text{(by Lemma 1)} \\
&> \sqrt{dN^2} && \text{(from Eqn. 1 and Eqn. 2)} \\
&= \texttt{dist}\big(\texttt{mid}^{-i}_{\mathbf{a}\oplus\mathbf{e}_i}, \texttt{minus}^{+i}_{\mathbf{a}}\big)
\end{aligned}$$

■

Since $\texttt{mid}^{-i}_{\mathbf{a}\oplus\mathbf{e}_i} \in \mathcal{P}'$ and $\textsc{Label}\big(\texttt{mid}^{-i}_{\mathbf{a}\oplus\mathbf{e}_i}\big) = \texttt{Bij}(\mathbf{a}\oplus\mathbf{e}_i) \neq \texttt{Bij}(\mathbf{a}) = \textsc{Label}\big(\texttt{minus}^{+i}_{\mathbf{a}}\big)$, it follows from Claim 6 that $\texttt{minus}^{+i}_{\mathbf{a}} \in \mathcal{P}'$. Therefore, we have $\big\{\texttt{plus}^{+i}_{\mathbf{a}}, \texttt{minus}^{+i}_{\mathbf{a}}\big\} \subseteq \big(\mathcal{P}' \cap \mathcal{P}(\mathbf{a})\big)$ which contradicts Eqn. 7.

Hence, $\mathcal{P}' \cap \textsc{Border}(\mathbf{a}) = \emptyset$ for each $\mathbf{a} \in V$ which concludes the proof of Lemma 2. ■

From Eqn. 6, Eqn. 7 and Lemma 2 it follows that

$$\forall \mathbf{a} \in V, \text{there is a point } \beta(\mathbf{a}) \in \textsc{Internal}_3(\mathbf{a}) \text{ such that} \big(\mathcal{P}' \cap \mathcal{P}(\mathbf{a})\big) = \big\{\beta(a)\big\} \tag{8}$$

We now construct a satisfying assignment for the instance $\mathcal{I} = (V, D, C)$ of $d$-dimensional geometric $\geq$-CSP.

**Lemma 3** The instance $\mathcal{I} = (V, D, C)$ of $d$-dimensional geometric $\geq$-CSP is satisfiable.

**Proof** For each $\mathbf{a} \in V$ let

$$\gamma(\mathbf{a}) = \beta(\mathbf{a}) \ominus C^* \cdot (\mathbf{a} - 1^d) = \beta(a) \ominus \texttt{origin}(\mathbf{a}) \oplus 1^d \tag{9}$$

We claim that the function $f : V \to D$ given by $f(\mathbf{a})[i] = \texttt{flip}_{\mathbf{a}[i]}\big(\gamma(\mathbf{a})[i]\big)$ for each $i \in [d]$ and each $\mathbf{a} \in V$ is a satisfying assignment for the instance $\mathcal{I} = (V, D, C)$ of $d$-dimensional geometric $\geq$-CSP. First we show that $f$ satisfies each unary constraint. For $\mathbf{a} \in V$ we have

$\beta(\mathbf{a}) \in \textsc{Internal}_3(\mathbf{a})$ (from Eqn. 8)
$\Rightarrow \gamma(\mathbf{a}) \in \textsc{Internal}_2(\mathbf{a})$ (from Section 2.1.3 & Eqn. 9)
$\Rightarrow f(\mathbf{a}) \in \textsc{Internal}_1(\mathbf{a})$ (from Section 2.1.2 & Observation 1)
$\Rightarrow f(\mathbf{a}) \in R_{\mathbf{a}}$ (from Section 2.1.1)

Next we show that $f$ satisfies each binary constraint. By Definition 3, every binary constraint in $C$ has the following structure: there exists a variable $\mathbf{b} \in V$ and an index $i \in [d]$ such that the binary constraint is $\big\langle (\mathbf{b}, \mathbf{b} \oplus \mathbf{e}_i), R_{\mathbf{b},\mathbf{b}\oplus\mathbf{e}_i} \big\rangle$ where $R_{\mathbf{b},\mathbf{b}\oplus\mathbf{e}_i} = \Big\{ \big(\mathbf{x}, \mathbf{y}\big) \subseteq R_{\mathbf{b}} \times R_{\mathbf{b}\oplus\mathbf{e}_i} \mid \mathbf{x}[i] \geq \mathbf{y}[i] \Big\}$. Hence, if we show that $f(\mathbf{b})[i] \geq f(\mathbf{b} \oplus \mathbf{e}_i)[i]$ then the binary constraint $\big\langle (\mathbf{b}, \mathbf{b} \oplus \mathbf{e}_i), R_{\mathbf{b},\mathbf{b}\oplus\mathbf{e}_i} \big\rangle$ is satisfied. In the remainder of this proof we denote $\mathbf{b} \oplus \mathbf{e}_i$ by $\mathbf{b}'$. There are two cases to consider depending on the parity of $\mathbf{b}[i]$:

1. $\mathbf{b}[i]$ is odd: By the definition of a consistent subset, we have

$$\texttt{dist}\big(\texttt{mid}_{\mathbf{b}}^{+i}, \beta(\mathbf{b})\big) < \texttt{dist}\big(\texttt{mid}_{\mathbf{b}}^{+i}, \beta(\mathbf{b} \oplus \mathbf{e}_i)\big) \tag{10}$$

Recall from Section 2.1.4 that $\texttt{mid}_{\mathbf{b}}^{+i} := \texttt{origin}(\mathbf{b}) \oplus \mathbf{e}_i \cdot \big((N-1) + (D^* - \epsilon)\big)$. Eqn. 4 implies that $\texttt{origin}(\mathbf{b} \oplus \mathbf{e}_i) = \texttt{origin}(\mathbf{b}) \oplus \mathbf{e}_i \cdot (N - 1 + 2D^*)$. Hence, Eqn. 9 and Eqn. 10 together imply that

$$\begin{aligned} &\Big(\big(N - \gamma(\mathbf{b})[i]\big) + (D^* - \epsilon)\Big)^2 + \sum_{j=1, j\neq i}^{d} \big(\gamma(\mathbf{b})[j] - 1\big)^2 \\ &< \Big(\big(\gamma(\mathbf{b}')[i] - 1\big) + (D^* + \epsilon)\Big)^2 + \sum_{j=1, j\neq i}^{d} \big(\gamma(\mathbf{b}')[j] - 1\big)^2 \end{aligned} \tag{11}$$

Since $\mathbf{b}[i]$ is odd, by Eqn. 3 we have $f(\mathbf{b})[i] = \gamma(\mathbf{b})[i]$ and $f(\mathbf{b}')[i] = (N+1) - \gamma(\mathbf{b}')[i]$. Eqn. 11 can be rewritten as

$$\begin{aligned} &\Big(\big(N - f(\mathbf{b})[i]\big) + (D^* - \epsilon)\Big)^2 + \sum_{j=1, j\neq i}^{d} \big(\gamma(\mathbf{b})[j] - 1\big)^2 \\ &< \Big(\big(N - f(\mathbf{b}')[i]\big) + (D^* + \epsilon)\Big)^2 + \sum_{j=1, j\neq i}^{d} \big(\gamma(\mathbf{b}')[j] - 1\big)^2 \end{aligned} \tag{12}$$

We now claim that $f(\mathbf{b})[i] \geq f(\mathbf{b}')[i]$. Suppose not, and $f(\mathbf{b})[i] < f(\mathbf{b}')[i]$. Since $f(\mathbf{b})[i], f(\mathbf{b}')[i] \in [N]$ it follows that $\big(f(\mathbf{b}')[i] - f(\mathbf{b})[i]\big) \geq 1$. Then from Eqn. 12 we have

$$\begin{aligned}
&(2d-2)N^2 = (d-1)N^2 + (d-1)N^2 \\
&\geq \sum_{j=1, j\neq i}^{d} \big(\gamma(\mathbf{b}')[j] - 1\big)^2 - \sum_{j=1, j\neq i}^{d} \big(\gamma(\mathbf{b})[j] - 1\big)^2 \qquad \text{(since } \gamma(\mathbf{b})[j], \gamma_{\mathbf{b}'}[j] \in [N] \;\forall\, j \in [d]\text{)} \\
&> \Big((N - f(\mathbf{b})[i]) + (D^* - \epsilon)\Big)^2 - \Big((N - f(\mathbf{b}')[i]) + (D^* + \epsilon)\Big)^2 \\
&= \Big(2D^* + (N - f(\mathbf{b})[i]) + (N - f(\mathbf{b}')[i])\Big) \times \big(f(\mathbf{b}')[i] - f(\mathbf{b})[i] - 2\epsilon\big) \\
&\geq 2D^* \times \frac{1}{2} = D^*
\end{aligned}$$

which is a contradiction since $D^* = 2d \cdot N^2$ (Eqn. 1). To obtain the penultimate inequality, we have used the bounds $f(\mathbf{b})[i], f(\mathbf{b}')[i] \leq N$ and $\big(f(\mathbf{b}')[i] - f(\mathbf{b})[i] - 2\epsilon\big) \geq 1/2$ since $\epsilon = 1/4$ (Eqn. 1).

2. $\mathbf{b}[i]$ is even: By the definition of a consistent subset, we have

$$\texttt{dist}\big(\texttt{mid}_{\mathbf{b}'}^{-i}, \beta(\mathbf{b}')\big) < \texttt{dist}\big(\texttt{mid}_{\mathbf{b}'}^{-i}, \beta(\mathbf{b})\big) \tag{13}$$

Recall from Section 2.1.4 that $\texttt{mid}_{\mathbf{b}'}^{-i} := \texttt{origin}(\mathbf{b}') \ominus \mathbf{e}_i \cdot (D^* - \epsilon)$. Eqn. 4 implies that $\texttt{origin}(\mathbf{b}) = \texttt{origin}(\mathbf{b}') \ominus \mathbf{e}_i \cdot (N - 1 + 2D^*)$. Hence, Eqn. 9 and Eqn. 13 together imply that

$$\begin{aligned}
&\Big((\gamma(\mathbf{b}')[i] - 1) + (D^* - \epsilon)\Big)^2 + \sum_{j=1, j\neq i}^{d} \big(\gamma(\mathbf{b}')[j] - 1\big)^2 \\
&< \Big((N - \gamma(\mathbf{b})[i]) + (D^* + \epsilon)\Big)^2 + \sum_{j=1, j\neq i}^{d} \big(\gamma(\mathbf{b})[j] - 1\big)^2
\end{aligned} \tag{14}$$

Since $\mathbf{b}[i]$ is even, by Eqn. 3 we have $f(\mathbf{b})[i] = N + 1 - \gamma(\mathbf{b})[i]$ and $f(\mathbf{b}')[i] = \gamma(\mathbf{b}')[i]$. Eqn. 14 can be rewritten as

$$\begin{aligned}
&\Big((f(\mathbf{b}')[i] - 1) + (D^* - \epsilon)\Big)^2 + \sum_{j=1, j\neq i}^{d} \big(\gamma(\mathbf{b}')[j] - 1\big)^2 \\
&< \Big((f(\mathbf{b})[i] - 1) + (D^* + \epsilon)\Big)^2 + \sum_{j=1, j\neq i}^{d} \big(\gamma(\mathbf{b})[j] - 1\big)^2
\end{aligned} \tag{15}$$

We now claim that $f(\mathbf{b})[i] \geq f(\mathbf{b}')[i]$. Suppose not, and $f(\mathbf{b})[i] < f(\mathbf{b}')[i]$. Since $f(\mathbf{b})[i], f(\mathbf{b}')[i] \in [N]$ it follows that $\big(f(\mathbf{b}')[i] - f(\mathbf{b})[i]\big) \geq 1$. Then from Eqn. 15 we have

$$\begin{aligned}
&(2d-2)N^2 = (d-1)N^2 + (d-1)N^2 \\
&\geq \sum_{j=1, j\neq i}^{d} \big(\gamma(\mathbf{b})[j] - 1\big)^2 - \sum_{j=1, j\neq i}^{d} \big(\gamma(\mathbf{b}')[j] - 1\big)^2 \quad \text{(since } \gamma(\mathbf{b})[j], \gamma_{\mathbf{b}'}[j] \in [N] \text{ for each } j \in [d]\text{)} \\
&> \Big((f(\mathbf{b}')[i] - 1) + (D^* - \epsilon)\Big)^2 - \Big((f(\mathbf{b})[i] - 1) + (D^* + \epsilon)\Big)^2 \\
&= \Big(2D^* + (f(\mathbf{b})[i] - 1) + (f(\mathbf{b}')[i] - 1)\Big) \times \big(f(\mathbf{b}')[i] - f(\mathbf{b})[i] - 2\epsilon\big) \\
&\geq 2D^* \times \frac{1}{2} = D^*
\end{aligned}$$

which is a contradiction since $D^* = 2d \cdot N^2$ (Eqn. 1). To obtain the penultimate inequality, we have used the bounds $f(\mathbf{b})[i], f(\mathbf{b}')[i] \geq 1$ and $\big(f(\mathbf{b}')[i] - f(\mathbf{b})[i] - 2\epsilon\big) \geq 1/2$ since $\epsilon = 1/4$ (Eqn. 1).

This concludes the proof of Lemma 3. ■

### 2.3. $\mathcal{I}$ has a satisfying assignment $\Rightarrow \mathcal{U}$ has a consistent subset of size $|V|$

Suppose that the instance $\mathcal{I} = (V, D, C)$ of $d$-dimensional geometric $\geq$-CSP has a satisfying assignment $f : V \to D$. We will now show that the instance $\mathcal{U} = (\mathcal{P}, \textsc{Label})$ of $k$-NNC-$(\mathbb{R}^d, \ell_2)$ has a consistent subset of size $|V|$.

Since $f : V \to D$ is a satisfying assignment for $\mathcal{I}$,

$$\text{for each } \mathbf{a} \in V, \text{ we have } f(\mathbf{a}) \in R_{\mathbf{a}} \tag{16}$$

$$\forall \mathbf{a} \in V, \forall i \in [d] \text{ such that } \mathbf{a} \,\&\, (\mathbf{a} \oplus \mathbf{e}_i) \text{ form an edge in} G_{\mathcal{I}}, \text{ we have} f(\mathbf{a})[i] \geq f(\mathbf{a} \oplus \mathbf{e}_i)[i] \tag{17}$$

We now construct a set $\mathcal{P}''$ (which we later show is a consistent subset in Lemma 5). For each $\mathbf{a} \in V$ let

$$g(\mathbf{a}) \in \mathbb{R}^d \text{ such that } g(\mathbf{a})[i] = \texttt{flip}_{\mathbf{a}[i]}\big(f(\mathbf{a})[i]\big) \tag{18}$$

$$h(\mathbf{a}) = \big(\texttt{origin}(\mathbf{a}) \ominus 1^d\big) \oplus g(\mathbf{a}) = C^* \cdot (\mathbf{a} \ominus 1^d) \oplus g(\mathbf{a}) \tag{19}$$

Define $\mathcal{P}'' = \big\{h(\mathbf{a}) \mid \mathbf{a} \in V\big\}$. Note that $|\mathcal{P}''| = |V|$.

**Lemma 4** $\big|\mathcal{P}'' \cap \textsc{Internal}_3(\mathbf{a})\big| = 1$ *for each* $\mathbf{a} \in V$

**Proof** We prove the lemma by showing that $h(\mathbf{a}) \in \textsc{Internal}_3(\mathbf{a})$ for each $\mathbf{a} \in V$. Fix any $\mathbf{b} \in V$. Then we have

$f(\mathbf{b}) \in R_{\mathbf{b}}$ (from Eqn. 16)
$\Rightarrow f(\mathbf{b}) \in \textsc{Internal}_1(\mathbf{b})$ (from Section 2.1.1)
$\Rightarrow g(\mathbf{b}) \in \textsc{Internal}_2(\mathbf{b})$ (from Section 2.1.2 and Eqn. 18)
$\Rightarrow h(\mathbf{b}) \in \textsc{Internal}_3(\mathbf{b})$ (from Section 2.1.3 and Eqn. 19)

■

We need a preliminary claim which gives a lower bound on the distance between internal points corresponding to different variables.

**Claim 7** Let $\mathbf{a}, \mathbf{a}' \in V$ such that $\mathbf{a} \neq \mathbf{a}'$. For any $\mathbf{q} \in \textsc{Internal}_3(\mathbf{a})$ and any $\mathbf{s} \in \textsc{Internal}_3(\mathbf{a}')$ we have $\texttt{dist}(\mathbf{q}, \mathbf{s}) \geq 2D^*$.

**Proof** Since $\mathbf{a} \neq \mathbf{a}'$ there exists some $j \in [d]$ such that $\mathbf{a}[j] \neq \mathbf{a}'[j]$. Let $\mathbf{q} = \texttt{origin}(\mathbf{a}) \oplus \mathbf{q}'$ and $\mathbf{s} = \texttt{origin}(\mathbf{a}') \oplus \mathbf{s}'$. Then it follows that for each $i \in [d]$ we have $0 \leq \mathbf{q}'[i], \mathbf{s}'[i] \leq (N-1)$.

$$\begin{aligned}
&\texttt{dist}\big(\mathbf{q}, \mathbf{s}\big) \\
&= \texttt{dist}\big(\texttt{origin}(\mathbf{a}) \oplus \mathbf{q}', \texttt{origin}(\mathbf{a}') \oplus \mathbf{s}'\big) \\
&\geq \Big|C^* \cdot \big(\mathbf{a}[j] - \mathbf{a}'[j]\big) + \big(\mathbf{q}'[j] - \mathbf{s}'[j]\big)\Big| && \text{(only counting along } j^{\text{th}}\text{-coordinate)} \\
&\geq \Big|C^* \cdot \big(\mathbf{a}[j] - \mathbf{a}'[j]\big)\Big| - \Big|\big(\mathbf{q}'[j] - \mathbf{s}'[j]\big)\Big| && \text{(by triangle inequality)} \\
&\geq C^* - (N-1) && \text{(since } \mathbf{a}[j] \neq \mathbf{a}'[j] \text{ and } 0 \leq \mathbf{q}'[j], \mathbf{s}'[j] \leq (N-1)\text{)} \\
&= 2D^* && \text{(from Eqn. 1)}
\end{aligned}$$

■

Finally, we are ready to show that

**Lemma 5** $\mathcal{P}''$ is a consistent subset for the instance $(\mathcal{P}, \textsc{Label})$ of $k$-NNC-$(\mathbb{R}^d, \ell_2)$.

**Proof** Fix any variable $\mathbf{a} \in V$. For any point $\mathbf{p} \in \mathcal{P}(\mathbf{a})$ and any $\mathbf{a}' \in V$ such that $\mathbf{a}' \neq \mathbf{a}$, we will now show that $\texttt{dist}\big(\mathbf{p}, h(\mathbf{a})\big) < \texttt{dist}\big(\mathbf{p}, h(\mathbf{a}')\big)$. Let $h(\mathbf{a}) = \texttt{origin}(\mathbf{a}) + \mathbf{t}$ and $h(\mathbf{a}') = \texttt{origin}(\mathbf{a}') + \mathbf{t}'$. By Eqn. 19, we have that

$$\mathbf{t} = g(\mathbf{a}) \ominus 1^d \quad \text{and} \quad \mathbf{t}' = g(\mathbf{a}') \ominus 1^d \tag{20}$$

If $\mathbf{p} \in \textsc{Internal}_3(\mathbf{a})$, then $(N-1) \geq \Big(\big|\mathbf{p}[\ell] - h(\mathbf{a})[\ell]\big|\Big) \geq 0$ for each $\ell \in [d]$. Hence, from Eqn. 1 and Eqn. 2 it follows that $\texttt{dist}\big(\mathbf{p}, h(\mathbf{a})\big) \leq \sqrt{dN^2} \leq D^*$. Since $\mathbf{a} \neq \mathbf{a}', \mathbf{p} \in \textsc{Internal}_3(\mathbf{a})$ and $h(\mathbf{a}') \in \textsc{Internal}_3(\mathbf{a}')$, from Claim 7 we have $\texttt{dist}\big(\mathbf{p}, h(\mathbf{a}')\big) \geq 2D^*$. Hence, $\texttt{dist}\big(\mathbf{p}, h(\mathbf{a}')\big) \geq 2D^* > D^* = \texttt{dist}\big(\mathbf{p}, h(\mathbf{a})\big)$.

Henceforth, we assume that $\mathbf{p} \in \textsc{Border}_3(\mathbf{a})$. By Eqn. 5, there exists $i \in [d]$ such that $\mathbf{p} \in \big(\textsc{Border}_{+i}(\mathbf{a}) \cup \textsc{Border}_{-i}(\mathbf{a})\big)$. We argue the case when $\mathbf{p} \in \textsc{Border}_{+i}(\mathbf{a})$: the case when $\mathbf{p} \in \textsc{Border}_{-i}(\mathbf{a})$ is analogous. There are two cases to consider depending on the parity of $\mathbf{a}[i]$:

1. $\underline{\mathbf{a}[i] \text{ is odd:}}$ By Section 2.1.4, it follows that $\mathbf{p} = \texttt{mid}_{\mathbf{a}}^{+i} = \texttt{origin}(\mathbf{a}) \oplus \mathbf{e}_i \cdot \big((N-1) + (D^* - \epsilon)\big)$. Then from Eqn. 1 we have

$$\texttt{dist}\big((\mathbf{p}, h(\mathbf{a})\big) = \sqrt{\Big((D^* - \epsilon) + (N-1) - t[i]\Big)^2 + \sum_{j=1, j\neq i}^{d} t[j]^2} \leq \sqrt{(D^* + N)^2 + dN^2} < 2D^* \tag{21}$$

where we have used Eqn. 1, Eqn. 2 and the fact that $0 \leq \mathbf{t}[\ell] \leq (N-1)$ for each $\ell \in [d]$.
If $\exists\, j \in [d]$ such that $j \neq i$ and $\mathbf{a}[j] \neq \mathbf{a}'[j]$, then

$$\begin{aligned}
&\texttt{dist}\big(\mathbf{p}, h(\mathbf{a}')\big) = \texttt{dist}\big(\texttt{mid}_{\mathbf{a}}^{+i}, \texttt{origin}(\mathbf{a}') \oplus \mathbf{t}'\big) \\
&\geq \Big|C^* \cdot \big(\mathbf{a}'[j] - \mathbf{a}[j]\big) + \mathbf{t}'[j]\Big| && \text{(only counting along } j^{\text{th}}\text{-coordinate)} \\
&\geq \Big|C^* \cdot \big(\mathbf{a}[j] - \mathbf{a}'[j]\big)\Big| - \Big|\mathbf{t}'[j]\Big| && \text{(by triangle inequality)} \\
&\geq C^* - (N-1) && \text{(since } \mathbf{a}[j] \neq \mathbf{a}'[j] \text{ and } 0 \leq \mathbf{t}'[j] \leq (N-1)\text{)} \\
&= 2D^* && \text{(from Eqn. 1)} \\
&> \texttt{dist}\big((\mathbf{p}, h(\mathbf{a})\big) && \text{(from Eqn. 21)}
\end{aligned}$$

Hence, we can assume that $\mathbf{a}[j] = \mathbf{a}'[j]$ for each $j \in [d]$ such that $j \neq i$. We have three subcases now:

- $\underline{\mathbf{a}'[i] \leq \mathbf{a}[i] - 1}$: In this subcase we have

$$\begin{aligned}
&\texttt{dist}\big(\mathbf{p}, h(\mathbf{a}')\big) \\
&\geq C^* \cdot \big(\mathbf{a}[i] - \mathbf{a}'[i]\big) + \big((D^* - \epsilon) + (N-1) - \mathbf{t}'[i]\big) && \text{(only counting along } i^{\text{th}}\text{-coordinate)} \\
&\geq C^* + \big(D^* - \epsilon\big) && \text{(since } 0 \leq \mathbf{t}'[i] \leq N-1 \text{ and } \big(\mathbf{a}[i] - \mathbf{a}'[i]\big) \geq 1\text{)} \\
&\geq 2D^* && \text{(from Eqn. 1 and Eqn. 2)} \\
&> \texttt{dist}\big((\mathbf{p}, h(\mathbf{a})\big) && \text{(from Eqn. 21)}
\end{aligned}$$

- $\mathbf{a}'[i] \geq \mathbf{a}[i]+2$: In this subcase we have

$$\begin{aligned}
&\texttt{dist}\big(\mathbf{p}, h(\mathbf{a}')\big) \\
&\geq C^* \cdot \big(\mathbf{a}'[i] - \mathbf{a}[i]\big) + \big(\mathbf{t}'[i] - (D^* - \epsilon) - (N-1)\big) && \text{(only counting along } i^{\text{th}}\text{-coordinate)} \\
&\geq 2C^* - \big(D^* + N\big) && \text{(since } \big(\mathbf{a}'[i] - \mathbf{a}[i]\big) \geq 2 \text{ and } 0 \leq \mathbf{t}'[i] \text{ and } \epsilon = \tfrac{1}{4}\text{)} \\
&= 2D^* && \text{(from Eqn. 1 and Eqn. 2)} \\
&> \texttt{dist}\big((\mathbf{p}, h(\mathbf{a})\big) && \text{(from Eqn. 21)}
\end{aligned}$$

- $\mathbf{a}'[i] = \mathbf{a}[i]+1$: In the last remaining subcase we have $\mathbf{a}' = \mathbf{a} + \mathbf{e}_i$. Hence, by Eqn. 17 we have that $f(\mathbf{a})[i] \geq f(\mathbf{a}')[i]$. Since $\mathbf{a}[i]$ is odd, by Eqn. 18 and Eqn. 3 we have $g(\mathbf{a})[i] = \texttt{flip}_{\mathbf{a}[i]}\big(f(\mathbf{a})[i]\big) = f(\mathbf{a})[i]$. Since $\mathbf{a}'[i] = \mathbf{a}[i]+1$ is even, by Eqn. 18 and Eqn. 3 we have $g(\mathbf{a}')[i] = \texttt{flip}_{\mathbf{a}'[i]}\big(f(\mathbf{a}')[i]\big) = N + 1 - f(\mathbf{a}')[i]$. Therefore, $f(\mathbf{a})[i] \geq f(\mathbf{a}')[i]$ implies that $g(\mathbf{a})[i] \geq N + 1 - g(\mathbf{a}')[i]$. From Eqn. 20 we can conclude that $\mathbf{t}[i] \geq N + 1 - \mathbf{t}'[i]$.

$$\begin{aligned}
&\texttt{dist}\big(\mathbf{p}, h(\mathbf{a})\big)^2 = \big((D^* - \epsilon) + (N + 1 - \mathbf{t}[i])\big)^2 + \sum_{j=1, j\neq i}^{d} \mathbf{t}[j]^2 && \text{(since } \mathbf{p} = \texttt{mid}_{\mathbf{a}}^{+i}\text{)} \\
&\leq \big((D^* - \epsilon) + \mathbf{t}'[i]\big)^2 + \sum_{j=1, j\neq i}^{d} \mathbf{t}[j]^2 && \text{(since } \mathbf{t}[i] \geq N + 1 - \mathbf{t}'[i]\text{)} \\
&\leq \big((D^* - \epsilon) + \mathbf{t}'[i]\big)^2 + (d-1)N^2 && \text{(since } 0 \leq \mathbf{t}[\ell] \leq N-1 \text{ for each } \ell \in [d]\text{)} \\
&< \big((D^* + \epsilon) + \mathbf{t}'[i]\big)^2 && \text{(from Eqn. 1, Eqn. 2 and since } \mathbf{t}'[i] \geq 0\text{)} \\
&\leq \big((D^* + \epsilon) + \mathbf{t}'[i]\big)^2 + \sum_{j=1, j\neq i}^{d} \mathbf{t}'[j]^2 = \texttt{dist}\big(\mathbf{p}, h(\mathbf{a}')\big)^2
\end{aligned}$$

2. $\mathbf{a}[i]$ is even: By Section 2.1.4, it follows that $\mathbf{p} = \texttt{plus}_{\mathbf{a}}^{+i}$ or $\mathbf{p} = \texttt{minus}_{\mathbf{a}}^{+i}$. We argue the case when $\mathbf{p} = \texttt{plus}_{\mathbf{a}}^{+i}$: the case when $\mathbf{p} = \texttt{minus}_{\mathbf{a}}^{+i}$ is analogous. From Section 2.1.4 and Eqn. 4 we have that $\mathbf{p} = \texttt{plus}_{\mathbf{a}}^{+i} = \texttt{origin}(\mathbf{a}) \oplus \mathbf{e}_i \cdot \big((N-1) + (D^* + \epsilon - N)\big) \oplus (1^d \ominus \mathbf{e}_i) \cdot N$. Eqn. 1 implies that

$$\texttt{dist}\big((\mathbf{p}, h(\mathbf{a})\big) = \sqrt{\big((D^* + \epsilon) - 1 - t[i]\big)^2 + \sum_{j=1, j\neq i}^{d} \big(N - t[j]\big)^2} \leq \sqrt{(D^*)^2 + dN^2} < 2D^* - N \tag{22}$$

where we have used Eqn. 1, Eqn. 2 and the fact that $0 \leq \mathbf{t}[j] \leq N-1$ for each $j \in [d]$. If $\exists\, j \in [d]$ such that $j \neq i$ and $\mathbf{a}[j] \neq \mathbf{a}'[j]$, then

$$\begin{aligned}
&\texttt{dist}\big(\mathbf{p}, h(\mathbf{a}')\big) = \texttt{dist}\big(\texttt{plus}_{\mathbf{a}}^{+i}, \texttt{origin}(\mathbf{a}') \oplus \mathbf{t}'\big) \\
&\geq \Big|C^* \cdot \big(\mathbf{a}'[j] - \mathbf{a}[j]\big) + \big(\mathbf{t}'[j] - N\big)\Big| && \text{(only counting along } j^{\text{th}}\text{-coordinate)} \\
&\geq \Big|C^* \cdot \big(\mathbf{a}[j] - \mathbf{a}'[j]\big)\Big| - \Big|\big(\mathbf{t}'[j] - N\big)\Big| && \text{(by triangle inequality)} \\
&\geq C^* - N && \text{(since } \mathbf{a}[j] \neq \mathbf{a}'[j] \text{ and } 0 \leq \mathbf{t}'[j] \leq (N-1)\text{)} \\
&> 2D^* - N && \text{(from Eqn. 1 and Eqn. 2)} \\
&= \texttt{dist}\big((\mathbf{p}, h(\mathbf{a})\big) && \text{(from Eqn. 22)}
\end{aligned}$$

Hence, we can assume that $\mathbf{a}[j] = \mathbf{a}'[j]$ for each $j \in [d]$ such that $j \neq i$. We have three subcases now:

- $\mathbf{a}'[i] \leq \mathbf{a}[i] - 1$: In this subcase we have

$$\begin{aligned}
&\texttt{dist}\big(\mathbf{p}, h(\mathbf{a}')\big) \\
&\geq C^* \cdot \big(\mathbf{a}[i] - \mathbf{a}'[i]\big) + \big((D^* + \epsilon - 1) - \mathbf{t}'[i]\big) && \text{(only counting along } i^{\text{th}}\text{-coordinate)} \\
&\geq C^* + \big(D^* + \epsilon - N\big) && \text{(since } 0 \leq \mathbf{t}'[i] \leq N-1 \text{ and } \big(\mathbf{a}[i] - \mathbf{a}'[i]\big) \geq 1\text{)} \\
&> 2D^* - N && \text{(from Eqn. 1 and Eqn. 2)} \\
&> \texttt{dist}\big((\mathbf{p}, h(\mathbf{a})\big) && \text{(from Eqn. 22)}
\end{aligned}$$

- $\mathbf{a}'[i] \geq \mathbf{a}[i] + 2$: In this subcase we have

$$\begin{aligned}
&\texttt{dist}\big(\mathbf{p}, h(\mathbf{a}')\big) \\
&\geq C^* \cdot \big(\mathbf{a}'[i] - \mathbf{a}[i]\big) + \big(\mathbf{t}'[i] - (D^* + \epsilon - 1)\big) && \text{(only counting along } i^{\text{th}}\text{-coordinate)} \\
&\geq 2C^* - (D^* + \epsilon) && \text{(since } \big(\mathbf{a}'[i] - \mathbf{a}[i]\big) \geq 2 \text{ and } \mathbf{t}'[i] \geq 0\text{)} \\
&> 2D^* - N && \text{(from Eqn. 1 and Eqn. 2)} \\
&> \texttt{dist}\big((\mathbf{p}, h(\mathbf{a})\big) && \text{(from Eqn. 22)}
\end{aligned}$$

- $\mathbf{a}'[i] = \mathbf{a}[i] + 1$: The last remaining subcase is $\mathbf{a}' = \mathbf{a} + \mathbf{e}_i$. Eqn. 4 implies that $\texttt{origin}(\mathbf{a}') = \texttt{origin}(\mathbf{a}) \oplus \mathbf{e}_i \cdot \big((N-1) + 2D^*\big)$. Hence we have

$$\begin{aligned}
&\texttt{dist}\big(\mathbf{p}, h(\mathbf{a})\big)^2 \\
&= \big((D^* + \epsilon) - 1 - \mathbf{t}[i]\big)^2 + \sum_{j=1, j\neq i}^{d} \big(N - \mathbf{t}[j]\big)^2 && \text{(since } \mathbf{p} = \texttt{plus}_{\mathbf{a}}^{+i}\text{)} \\
&< (D^*)^2 + d \cdot N^2 && \text{(since } \epsilon = \tfrac{1}{4} \text{ and } 0 \leq \mathbf{t}'[\ell] \leq N-1 \text{ for each } \ell \in [d]\text{)} \\
&\leq (D^* + N - \epsilon)^2 && \text{(from Eqn. 1 and Eqn. 2)} \\
&\leq \big(D^* - \epsilon + N + \mathbf{t}'[i]\big)^2 && \text{(since } \mathbf{t}'[i] \geq 0\text{)} \\
&\leq \big(D^* - \epsilon + N + \mathbf{t}'[i]\big)^2 + \sum_{j=1, j\neq i}^{d} \big(N - \mathbf{t}'[j]\big)^2 \\
&= \texttt{dist}\big(\mathbf{p}, h(\mathbf{a}')\big)^2
\end{aligned}$$

This concludes the proof of Lemma 5. ■

### 2.4. Proofs of Theorem 1 and Theorem 2

Finally, we are now ready to prove Theorem 1 and Theorem 2:

**Proof** Given an instance $\mathcal{I} = (V, D, C)$ of a $d$-dimensional geometric $\geq$-CSP, we designed a reduction (Section 2.1) to build in $|\mathcal{I}|^{O(1)}$ time an instance $(\mathcal{P}, \text{LABEL})$ of $k$-NNC-$(\mathbb{R}^d, \ell_2)$ such that $k = |V|$ and $|\mathcal{P}| := n \leq |C| + 3 \cdot |V|^2 = |\mathcal{I}|^{O(1)}$. The correctness of this reduction follows from the two directions shown in Section 2.2 and Section 2.3: $\mathcal{I} = (V, D, C)$ has a satisfying assignment if and only if $(\mathcal{P}, \text{LABEL})$ has a consistent subset of size $|V|$. Since the $d$-dimensional geometric $\geq$-CSP problem is W[1]-hard parameterized by $|V| + d$ (Theorem 3), it follows that $k$-NNC-$(\mathbb{R}^d, \ell_2)$ is W[1]-hard parameterized by $k + d$, i.e., . This concludes the proof of Theorem 1.

Theorem 4 states that assuming the Exponential Time Hypothesis (ETH) there is no $d \geq 2$ and computable function $f$ such that instances $\mathcal{I} = (V, D, C)$ of $d$-dimensional geometric $\geq$-CSP can be solved in $f(|V|) \cdot |\mathcal{I}|^{o(|V|^{1-1/d})}$ time. Since our reduction (Section 2.1) converts an instance of $\mathcal{I} = (V, D, C)$ of $d$-dimensional geometric $\geq$-CSP in $|\mathcal{I}|^{O(1)}$ time into an equivalent instance $(\mathcal{P}, \text{LABEL})$ of $k$-NNC-$(\mathbb{R}^d, \ell_2)$ with $k = |V|$ and the number of points $n = |\mathcal{I}|^{O(1)}$. Hence, it follows that assuming the Exponential Time Hypothesis (ETH) there is no $d \geq 2$ and computable function $f$ such that $k$-NNC-$(\mathbb{R}^d, \ell_2)$ can be solved in $f(k) \cdot n^{o(k^{1-1/d})}$ time where $n$ is the number of balls and $k$ is the size of the consistent subset. This concludes the proof of Theorem 2. ■

We now describe the small changes needed to make the same construction work for $\ell_p$-metrics for $1 \leq p \leq \infty$:

**Remark 3** Our lower bound extends to the $\ell_\infty$-metric with exactly the same construction: in fact some of the proofs are simpler for $\ell_\infty$ as compared to $\ell_2$. The only minor change needed to make the lower bound work for the $\ell_q$-metric (for $q \in \mathbb{N}$) is to change the value of $D$ in Eqn. 1 to $2dN^q$ instead of $2dN^2$, and all the calculations go through.

## References


Fabrizio Angiulli. Fast condensed nearest neighbor rule. In *International Conference on Machine Learning (ICML)*, pages 25–32, 2005.

Sandip Banerjee, Sujoy Bhore, and Rajesh Chitnis. Algorithms and Hardness Results for Nearest Neighbor Problems in Bicolored Point Sets. In *Latin American Theoretical Informatics (LATIN)*, pages 80–93, 2018.

Ricardo Barandela, Francesc J. Ferri, and José Salvador Sánchez. Decision boundary preserving prototype selection for nearest neighbor classification. *Int. J. Pattern Recognit. Artif. Intell.*, 19 (6):787–806, 2005.

Ahmad Biniaz, Sergio Cabello, Paz Carmi, Jean-Lou De Carufel, Anil Maheshwari, Saeed Mehrabi, and Michiel H. M. Smid. On the Minimum Consistent Subset Problem. In *Algorithms and Data Structures Symposium (WADS)*, pages 155–167, 2019.

Kenneth L. Clarkson. An Algorithm for Approximate Closest-Point Queries. In *Symposium on Computational Geometry (SoCG)*, pages 160–164, 1994.

Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory*, 13(1):21–27, 1967.

Marek Cygan, Fedor V. Fomin, Lukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michal Pilipczuk, and Saket Saurabh. *Parameterized Algorithms*. Springer, 2015. ISBN 978-3-319-21274-6.

V. Susheela Devi and M. Narasimha Murty. An incremental prototype set building technique. *Pattern Recognit.*, 35(2):505–513, 2002.

Luc Devroye. On the inequality of Cover and Hart in nearest neighbor discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (1):75–78, 1981.

Evelyn Fix and Joseph L. Hodges. Discriminatory analysis, nonparametric discrimination: Consistency properties. *US Air Force School of Aviation Medicine, Technical Report*, 4:21–49, 1951.

Alejandro Flores-Velazco. Social Distancing is Good for Points too! In *Canadian Conference on Computational Geometry, CCCG*, pages 352–358, 2020.

Alejandro Flores-Velazco and David M. Mount. Guarantees on Nearest-Neighbor Condensation heuristics. In *Canadian Conference on Computational Geometry CCCG*, pages 87–93, 2019.

Salvador García, Joaquín Derrac, José Ramón Cano, and Francisco Herrera. Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(3):417–435, 2012.

Geoffrey W. Gates. The reduced nearest neighbor rule (corresp.). *IEEE Trans. Inf. Theory*, 18(3): 431–433, 1972.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Pinhas Nisnevitch. Near-optimal sample compression for nearest neighbors. In *Conference on Neural Information Processing Systems (NIPS)*, pages 370–378, 2014.

Peter E. Hart. The condensed nearest neighbor rule (corresp.). *IEEE Trans. Inf. Theory*, 14(3): 515–516, 1968.

Russell Impagliazzo and Ramamohan Paturi. On the Complexity of $k$-SAT. *J. Comput. Syst. Sci.*, 62(2):367–375, 2001.

Russell Impagliazzo, Ramamohan Paturi, and Francis Zane. Which Problems Have Strongly Exponential Complexity? *J. Comput. Syst. Sci.*, 63(4):512–530, 2001.

Norbert Jankowski and Marek Grochowski. Comparison of Instances Selection Algorithms I. Algorithms Survey. In *International Conference on Artificial Intelligence and Soft Computing (ICAISC)*, pages 598–603, 2004.

Kamyar Khodamoradi, Ramesh Krishnamurti, and Bodhayan Roy. Consistent Subset Problem with Two Labels. In *Annual International Conference on Algorithms and Discrete Applied Mathematics (CALDAM)*, pages 131–142, 2018.

Aryeh Kontorovich and Roi Weiss. A Bayes consistent 1-NN classifier. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38 of *JMLR Workshop and Conference Proceedings*. JMLR.org, 2015.

Robert Krauthgamer and James R. Lee. Navigating nets: simple algorithms for proximity search. In *Symposium on Discrete Algorithms (SODA)*, pages 798–807, 2004.

Daniel Lokshtanov, Dániel Marx, and Saket Saurabh. Lower bounds based on the exponential time hypothesis. *Bull. EATCS*, 105:41–72, 2011.

Dániel Marx and Anastasios Sidiropoulos. The limited blessing of low dimensionality: when $1-1/d$ is the best possible exponent for $d$-dimensional geometric problems. In *Symposium on Computational Geometry (SoCG)*, page 67, 2014.

G. L. Ritter, Hugh B. Woodruff, Stephen R. Lowry, and Thomas L. Isenhour. An algorithm for a selective nearest neighbor decision rule (corresp.). *IEEE Trans. Inf. Theory*, 21(6):665–669, 1975.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014. ISBN 978-1-10-705713-5.

Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.

Godfried T. Toussaint. Open problems in geometric methods for instance-based learning. In *Discrete and Computational Geometry, Japanese Conference, JCDCG 2002, Tokyo, Japan, December 6-9, 2002, Revised Papers*, pages 273–283, 2002.

Gordon T. Wilfong. Nearest Neighbor Problems. In *Symposium on Computational Geometry (SoCG)*, pages 224–233, 1991.

D. Randall Wilson and Tony R. Martinez. Reduction techniques for instance-based learning algorithms. *Mach. Learn.*, 38(3):257–286, 2000.

AV Zukhba. NP-completeness of the problem of prototype selection in the nearest neighbor method. *Pattern Recognition and Image Analysis*, 20(4):484–494, 2010.