New Coresets for Fair Clustering

Chansophea Wathanak In **Yi Li**
School of Physical and Mathematical Sciences
Nanyang Technological University
inch0002@e.ntu.edu.sg
yili@ntu.edu.sg

Xuan Wu
John Hopcroft Center for Computer Science
Shanghai Jiao Tong University
wuxuan2026@sjtu.edu.cn

Abstract

We construct the first coreset of size near-linear in k for fair k-MEDIANS in general metric spaces. Combined with the standard merge-and-reduce framework, our construction also gives the first streaming algorithm for fair k-MEDIANS with space complexity near-linear in k.

The main technical innovation is a distributional reinterpretation of capacitated clustering, in which the cost of assigning a dataset to a capacitated center set can be expressed as the earth mover's distance between two discrete distributions. This viewpoint enables us to use tools from optimal transport to analyse uniform sampling. A key ingredient in this analysis is a new family of problem-specific ϵ-nets for the earth mover's distance.

1 Introduction

We consider fair k-MEDIANS clustering in a general metric space, a notion proposed by Chierichetti et al. [16] to avoid discriminatory outputs of clustering algorithms. Roughly speaking, the fair k-MEDIANS problem requires each cluster to satisfy prescribed representation constraints for each group. Specifically, let (M, ρ) be a metric space, $X \subset M$, and suppose that $X = G_1 \cup \cdots \cup G_g$, where $G_1, \ldots, G_g$ are groups, possibly overlapping. Let $\alpha, \beta \in \mathbb{Z}_{\geq 0}^{k \times g}$. A k-clustering $\{C_1, \ldots, C_k\}$ of X is said to be (α, β)-*fair* if $\alpha_{ij} \leq |C_i \cap G_j| \leq \beta_{ij}$ for all $i \in [k]$ and $j \in [g]$. The (α, β)-fair k-MEDIANS problem asks to find an (α, β)-fair k-clustering such that $\sum_{j \in [k]} \sum_{x \in C_j} \rho(x, a_j)$ is minimized, where a_j is the center of cluster C_j.

The additional constraints imposed on the clustering objectives are well motivated by considerations of social fairness and have received considerable attention in the machine learning community [16, 9, 28, 18]. However, these fairness constraints also introduce significant computational challenges. In the general setting, k-MEDIANS is NP-hard, and it remains open whether it admits a polynomial-time constant-factor approximation algorithm [28]. To this end, coresets have been considered as an efficient data reduction technique to mitigate the computational difficulty. Intuitively, a coreset is a small weighted proxy of the dataset such that the clustering cost is approximately preserved for every possible set of centers. In the classical unconstrained setting, coresets have been studied extensively, and almost optimal constructions are now known in Euclidean and other metric spaces [25, 24, 15, 38, 21, 32, 20, 19, 17, 31].

Classical coresets for clustering are generally inadequate for fair clustering because the fairness constraints are not captured by the unconstrained clustering cost. To address this issue, a stronger notion, called an assignment-preserving coreset, was considered [28, 12]. Roughly speaking, an assignment-preserving coreset approximately preserves the clustering cost for every possible set of centers and every feasible capacity constraint.[1].

[1] Fairness constraints can be regarded as a special form of capacity constraints

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Through a series of works [44, 41, 28, 8, 30], the best known assignment-preserving coreset size for k-MEDIANS in an n-point metric space is $\tilde{O}(k^2\varepsilon^{-2}\log n)$. The quadratic dependence on k arises from the geometric decompositions used in the existing algorithms. Achieving a near-linear dependence on k is therefore a natural and important goal. First, $\Omega(k)$ is a lower bound [7, 19, 31]; second, the parameter k captures the structural complexity of the clustering problem: for a fixed k, varying ε usually changes the solution only mildly, whereas changing k could fundamentally alter the solution's structure.

Our Contributions In this paper, we construct the first assignment-preserving coreset of size near-linear in k for k-MEDIANS in general metric spaces. We also obtain the first coresets for fair k-MEDIANS in high-dimensional Euclidean spaces whose sizes are near-linear in k and independent of the ambient dimension. Combined with the standard merge-and-reduce framework (see, e.g. [1, 25, 15]), our construction further gives the first streaming algorithm for fair k-MEDIANS with space complexity near-linear in k.

Our approach is based on a distributional interpretation of the capacitated clustering cost. Namely, for a dataset X and a capacitated center set C with the same total capacity, we regard X and C as discrete distributions and interpret the clustering cost as proportional to the earth-mover's distance (EMD). This interpretation allows us to reduce the coreset construction to the analysis of empirical distributions under EMD and further enables us to employ tools from optimal transport. The main technical innovation is a new family of problem-specific ε-nets for EMD, which enables us to establish the coreset guarantee simultaneously for all sets of k capacitated centers.

1.1 Preliminaries

To state our results, we first introduce the necessary notation and formalize the distributional interpretation of capacitated clustering.

In a metric space (M, ρ), a ball $B(a, r)$ in (M, ρ) is the set $B(a, r) = \{x \in M \mid \rho(a, x) \le r\}$.

Definition 1.1 (Coreset for fair clustering [28])**.** An ε-coreset for fair k-MEDIANS is a subset $S \subseteq M$ with a weight function $w : S \to \mathbb{R}^k_{\ge 0}$ such that for every (α, β)-fair k-clustering $\{C_1, \dots, C_k\}$ with centers $\{a_1, \dots, a_k\}$,

$$(1-\varepsilon)\sum_{i\in[k],s\in S}(w(s))_i\rho(s,a_i) \le \sum_{i\in[k]}\sum_{x\in C_i}\rho(x,a_i) \le (1+\varepsilon)\sum_{i\in[k],s\in S}(w(s))_i\rho(s,a_i).$$

The fairness constraint can be interpreted as a special form of capacity constraint.

Definition 1.2 (Capacitated Clustering)**.** Let X denote a dataset in a metric space (M, ρ). Let $C \subset M$ denote a center set, and $w : C \to \mathbb{R}_{\ge 0}$ be a capacity function on C. That is, $\forall c \in C, w(c)$ denotes the maximum number of data points that can be connected to c in a clustering. If $\sum_{c\in C} w(c) = |X|$, we say X and C are *consistent* and define $\mathrm{cost}(X, C)$ to be the minimum cost maximum matching between X and C[2].

Dataset as Distribution To make our results as broadly applicable as possible and to simplify the presentation, we interpret datasets as distributions throughout the paper. This perspective has a significant advantage that the clustering cost between a dataset X and a capacitated center set C, provided they have the same total mass, is proportional to the earth mover's distance between the corresponding distributions.

Let p and q be two discrete probability distributions on M. A transportation matrix from p to q is a map $\Pi : \mathrm{supp}(p) \times \mathrm{supp}(q) \to \mathbb{R}_{\ge 0}$ which satisfies the marginal constraints:

$$\sum_{y\in\mathrm{supp}(q)}\Pi(x,y) = p(x),\ \forall x \in \mathrm{supp}(p) \quad \text{and} \quad \sum_{x\in\mathrm{supp}(p)}\Pi(x,y) = q(y),\ \forall y \in \mathrm{supp}(q).$$

The earth-mover's distance between p and q is defined as

$$\mathrm{EMD}_M(p,q) = \min_{\Pi\in\Gamma(p,q)}\sum_{\substack{x\in\mathrm{supp}(p)\\ y\in\mathrm{supp}(q)}}\Pi(x,y)\cdot\rho(x,y),$$

[2]Specifically, the cost of the edge between $x \in X$ and $c \in C$ is $\rho(x, c)$ and each $c \in C$ has multiplicity $w(c)$.

where $\Gamma(p, q)$ is the set of all transportation matrices from p to q. When the underlying metric space is clear from the context, we simply write EMD for EMD_M.

Fact 1. *If X is a dataset of n points and C is a weighted center set consistent with X, then $\frac{1}{n} \cdot \mathrm{cost}(X, C) = \mathrm{EMD}(\frac{1}{n} \cdot X, \frac{1}{n} \cdot C)$. Here, $\frac{1}{n} \cdot X$ denotes the empirical distribution p over X, i.e., $p(x) = \frac{1}{n}$ for all $x \in X$, and $\frac{1}{n} \cdot C$ denotes the distribution q supported on C with $q(x) = \frac{w(c)}{n}$ for all $c \in C$.*

Definition 1.3 (Subdistribution)**.** A distribution p' is a *subdistribution* of p if $\mathrm{supp}(p') \subseteq \mathrm{supp}(p)$.

For a subset $A \subset \mathrm{supp}(p)$, we let $p|_A$ denote the restriction of p on A, that is, $p|_A$ is a measure defined as $p|_A(x) = p(x)$ for all $x \in A$ and $p|_A(x) = 0$ for all $x \notin A$.

We are now ready to define assignment-preserving coresets.

Definition 1.4 (Assignment-preserving Coresets [12])**.** Let p be a distribution on a metric space (M, ρ). A subdistribution $\widehat{p}$ of p is said to be an assignment-preserving (ε, k, Δ)-*coreset* of p if, for every distribution q on M with $|\mathrm{supp}(q)| \leq k$, the following holds:

$$(1 - \varepsilon) \cdot \mathrm{EMD}(p, q) - \Delta \leq \mathrm{EMD}(\widehat{p}, q) \leq (1 + \varepsilon) \cdot \mathrm{EMD}(p, q) + \Delta.$$

When the additive error $\Delta = 0$, the subdistribution $\widehat{p}$ is also called an (ε, k)-*coreset*.

Fact 2 (Coresets are mergeable)**.** *If μ_1 is an $(\varepsilon, k, \Delta_1)$-coreset of p_1 and μ_2 is an $(\varepsilon, k, \Delta_2)$-coreset of p_2, and $\beta \in (0, 1)$, then $\beta\mu_1 + (1-\beta)\mu_2$ is an $(\varepsilon, k, \beta\Delta_1 + (1-\beta)\Delta_2)$-coreset of $\beta p_1 + (1-\beta)p_2$.*

1.2 Our Results

We state our main theorem in a very general form. We note that it is standard in the literature [28, 12, 30] that the construction of fair coresets can be reduced to the construction of assignment-preserving coresets.

Theorem 3. *There exists an algorithm which, given any distribution p supported on an n-point metric space (M, ρ), returns an assignment-preserving (ε, k)-coreset $\widehat{p}$ of p with $|\mathrm{supp}(\widehat{p})| \leq \tilde{O}(k\varepsilon^{-4} \log n \log \frac{\tau}{\varepsilon})$, where $\tau = \min(\frac{1}{p_{\min}}, \lambda)$, $p_{\min} = \min_{x \in \mathrm{supp}(p)} p(x)$ and λ denotes the aspect ratio[3] of $\mathrm{supp}(p)$.*

Compared with the previous $\tilde{O}(k^2\varepsilon^{-2} \log n)$ bound, this improves the dependence on k from quadratic to near-linear, at the cost of a larger dependence on ε^{-1} and an additional logarithmic factor.

We remark that in $\mathbb{R}^d$, the previous near-linear bounds in k for EMD coresets are $O(k\varepsilon^{-d})$ [44, 28] and are thus practical only in low dimensions. Our approach combines uniform sampling with a problem-specific net over sparse distributions and their transportation structure (Section 5.2), which yields a fixed polynomial dependence on ε^{-1}. In Euclidean spaces, terminal embeddings and iterative size reduction then remove the remaining dependence on d, giving the following corollary.

Corollary 4. *There is an algorithm which, given any discrete distribution p in $\mathbb{R}^d$ with $|\mathrm{supp}(p)| = n$, computes in time $\tilde{O}(nd)$ an assignment-preserving (ε, k)-coreset $\widehat{p}$ of p with constant probability and $|\mathrm{supp}(\widehat{p})| \leq \tilde{O}(k\varepsilon^{-6} \log n)$.*

We remark that the running time of the algorithm in Theorem 3 is determined by the time to compute a constant-factor approximation for k-MEDIANS in the underlying metric space. In Euclidean spaces, a near-linear time algorithm is available [37].

Combining our coresets and the standard merge-and-reduce framework (see, e.g. [1, 25, 15]), we obtain the first coreset for fair k-MEDIANS of size $k \cdot (\varepsilon^{-1} \log n)^{O(1)}$ and the first streaming algorithm[4] with space complexity $k \cdot (\varepsilon^{-1} \log \Delta)^{O(1)}$ in $\mathbb{R}^d$.

[3] The aspect ratio of a set A is ratio between the largest and the smallest distance between two different points of A.

[4] We note that the paper [18] contains an omission: the claim on page 2 that adapting [12] to the insertion-only stream model with the standard merge-and-reduce framework yields a space linear in k is inaccurate. A direct application of [12] incurs a k^3 dependence in space since the coreset size there is $\tilde{O}(k^3\varepsilon^{-6})$ and, to the best of our knowledge, no coresets of $o(k^2)$ size were known in high-dimensional Euclidean space ($d \gg \log k$) prior to our work. This omission has been confirmed in private communication with an author of [18].

Corollary 5. *Let $X \subseteq \mathbb{R}^d$ be a set of n points and let G be the number of distinct combinations of groups to which a single point may belong. Then, one can compute an universal ε-coreset of X for fair k-clustering, which is valid for all choices of (α, β), with size $\tilde{O}(Gk\varepsilon^{-6}\log n)$ in time $\tilde{O}(nd + Gk\varepsilon^{-6})$. Moreover, the coreset leads to a streaming algorithm for (α, β)-fair k-MEDIANS on an insertion-only data stream in $[\Delta]^d$, using $Gk \cdot (\varepsilon^{-1}\log\Delta)^{O(1)}$ words of space.*

2 Technical Overview

The state-of-the-art approaches [8, 12, 30] for constructing assignment-preserving coresets follow a common scheme: decompose the distribution p into $\tilde{O}(k)$ subdistributions with bounded support, and then show that the empirical distribution for each subdistribution forms a coreset (with an affordable additive error). However, this strategy seems to yield coresets of size at least $\Omega(k^2)$ (one factor of k from the decomposition and the other from a union bound). To achieve a better coreset size, we consider a more powerful regularity condition (Definition 4.1), motivated by the group sampling techniques in clustering [20, 31, 29].

Specifically, a distribution p is said to be (r, t)-regular if its support can be partitioned into t sets $B_1 \subseteq B(a_1, r), \ldots, B_t \subseteq B(a_t, r)$ of radius at most r such that $p(B_i) \in [\frac{1}{2t}, \frac{2}{t}]$ for every $i \in [t]$, that is, each part contains roughly the same mass. We apply this notion with $t \le k$. We show that an empirical distribution p_m of size $m = \tilde{O}(k\varepsilon^{-4}\log n)$, combined with a natural and necessary scaling step (Line 5 of Algorithm 1), constitutes an $(\varepsilon, k, \varepsilon r)$-coreset for p. Let μ be the output of Algorithm 1. To prove the correctness of μ, we need to show that for every k-sparse distribution q, there exists an optimal transportation matrix Π between μ and q whose cost is close to that of Λ, the optimal transportation between p and q.

Our design of Π relies on a classical fact from optimal transport (Lemma 11), which states that Π is optimal if $\mathrm{supp}(\Pi) \subset \mathrm{supp}(\Lambda)$. In other words, if a transportation matrix for the subdistribution does not introduce new point pairs, it is automatically optimal. This suggests a natural candidate

$$\Pi(x, y) = \mu(x) \cdot \Lambda(x, y)/p(x), \quad x \in \mathrm{supp}(\mu),\ y \in \mathrm{supp}(q).$$

While this choice ensures that $\mathrm{supp}(\Pi) \subset \mathrm{supp}(\Lambda)$, it is not guaranteed that Π is a transportation matrix. Since μ is random, the marginals $\sum_{x \in \mathrm{supp}(\mu)} \Pi(x, y)$ may not equal $q(y)$.

To overcome this issue, we make adjustments to the transportations in Π. Specifically, we show in Lemma 12 that, for a fixed k-sparse distribution q, with high probability over the random μ, we can modify μ into a new distribution w and modify Π into a true transportation matrix Θ : $\mathrm{supp}(w) \times \mathrm{supp}(q) \to \mathbb{R}_{\ge 0}$ such that $\mathrm{EMD}(\mu, q) \approx \mathrm{EMD}(w, q)$.

To manage the cost error incurred by these adjustments, we classify pairs (x, y), where $x \in B_i$ and $y \in \mathrm{supp}(q)$, into close pairs (where $\rho(a_i, y) \le \varepsilon^{-1} r$) and far pairs. Each close transportation has length at most $(1 + \varepsilon^{-1})r$, so changing $O(\varepsilon^2)$ total transportation mass on close pairs changes its cost by at most $O(\varepsilon r)$. For far pairs, we observe that points within a regularity part behave essentially like a single point: as long as the total mass sent from the part to a fixed far point is preserved, the cost changes by a factor of $1 \pm O(\varepsilon)$. Indeed, moving mass within the part changes each source-to-destination distance by at most $2r$, while every far transportation has length at least $(\varepsilon^{-1} - 1)r$. This motivates us to enforce that in each part B_i, the total mass in far transportation is kept unchanged after the adjustments. We use two L_1-type distances (see Lemma 9) to control the error arising from the adjustments made to the close and far pairs, giving $\mathrm{EMD}(\mu, q) \approx \mathrm{EMD}(w, q)$.

A similar cost analysis shows that with high probability, $\mathrm{EMD}(w, \nu) \approx \mathrm{EMD}(p, \nu)$, and thus $\mathrm{EMD}(\mu, \nu) \approx \mathrm{EMD}(p, \nu)$, for each fixed $O(k)$-sparse distribution ν. To extend such approximation to all k-sparse distributions q simultaneously (i.e. the coreset property of μ), we employ a net argument (Section 5.2). Unlike the L_1-distance, discretizing densities does not necessarily control the EMD, due to far transportations. To resolve this, we observe that any far transportation from $x \in B_i$ to $y \in \mathrm{supp}(q)$ can be rerouted through the ball center a_i (so the transportation $x \to y$ becomes $x \to a_i \to y$). Such rerouting creates a new auxiliary $2k$-sparse distribution $\widehat{q}$ supported on $\mathrm{supp}(q) \cup \{a_1, \ldots, a_t\}$ and clearly $|\mathrm{EMD}(\mu, q) - \mathrm{EMD}(\mu, \widehat{q})| \le \mathrm{EMD}(\widehat{q}, q)$. We show in the proof of Lemma 14 that it suffices to construct a discretization family (which we still refer to as a "net") for all such auxiliary distributions $\widehat{q}$.

We now turn to the construction of the net. By the construction of $\widehat{q}$, every point in $\mathrm{supp}(\widehat{q})$ is close to at least one center a_i ($i \in [t]$). To ensure that an L_1-net of $\widehat{q}$ is indeed a net for EMD, we need to take extra care with density discretization. Specifically, we build an auxiliary graph G on $\mathrm{supp}(q) \cup \{a_1, \ldots, a_t\}$, connecting a_i to $y \in \mathrm{supp}(q)$ whenever $\rho(a_i, y) \le \varepsilon^{-1} r$. Thus, G has at most $2k$ vertices and every connected component of G has diameter $O(k\varepsilon^{-1} r)$. The connected components of G induce a partition of $\mathrm{supp}(\widehat{q})$ and we discretize the densities component by component. This yields a net point $\tilde{q}$; by choosing the discretization precision appropriately, we can ensure $L_1(\widehat{q}, \tilde{q}) \le \varepsilon^2/k$. Since discretization perturbs mass only inside components and any two vertices in a component are at a distance $O(k\varepsilon^{-1} r)$, we obtain the bound $\mathrm{EMD}(\widehat{q}, \tilde{q}) \le O(k\varepsilon^{-1} r) \cdot L_1(\widehat{q}, \tilde{q}) \le \varepsilon r$.

3 Related Work

EMD Coresets in Euclidean Spaces The exponential dependence on d in the previous $O(k\varepsilon^{-d})$ bounds comes from discretizing the ambient space. The coreset used by [44], which is based on [25], rounds points to multiscale grids whose complexity is ε^{-d}. The construction of [28] instead projects the data onto rays determined by an ε-net of directions and builds a one-dimensional coreset on each ray; the number of directions is again ε^{-d}. Although the two constructions differ, both pay a factor exponential in d for an ambient geometric discretization.

Coresets for Robust Clustering Apart from fair clustering, coresets have also been constructed for other constrained clustering problems, such as robust clustering (e.g. k-MEDIANS with m outliers). The first robust coresets were constructed by Feldman et al. [22], with a size that depends exponentially on the number of outliers. Huang et al. [27] later constructed the first robust coresets of fully polynomial size. In a series of follow-up works [31, 33, 35, 29], the coreset size for robust clustering has been improved to be nearly optimal. Very recently, it was proved in [29] that in a metric space with a bounded VC or doubling dimension, the optimal coreset size for k-MEDIANS with m outliers is $m + Q$ (up to polylogarithmic factors), where Q denotes the optimal coreset size under the unconstrained setting.

Computing EMD in $\mathbb{R}^d$ Approximation algorithms for discrete EMD in $\mathbb{R}^d$ have received considerable attention from the theoretical computer science community. Early approaches [14, 34, 4] focused on embedding EMD into metrics that are easier to compute, such as the ℓ_1-metric or tree metric, which often incur a logarithmic approximation factor. To obtain a constant-factor or $(1+\varepsilon)$-approximation, a standard approach is to solve the min-cost flow problem in a geometric spanner [42, 23, 36, 5], which is a sparse graph that approximates all pairwise distances. Very recently, sub-quadratic algorithms beyond spanner approaches have been proposed [11, 10].

Another related line of work studies the semi-discrete setting, where one distribution is absolutely continuous and the other one has a finite support. Existing algorithms typically exploit the structure of the optimal solution, which is known to be induced by a weighted Voronoi tessellation, and compute this tessellation using numerical optimization; see, e.g. [6, 39, 26]. More recently, combinatorial ideas have been applied to this problem, notably in [3, 2].

4 Proof of Theorem 3

Definition 4.1 ((r,t)-Regular Distribution)**.** A distribution p is said to be (r,t)-*regular* if $\mathrm{supp}(p)$ can be partitioned into t disjoint sets $\{B_i\}_{i=1}^t$ such that the following conditions hold for every $i \in [t]$:

(i) each B_i has radius at most r, that is, $B_i \subseteq B(a_i, r)$ for some center $a_i \in M$, and

(ii) $\frac{1}{2t} \le p(B_i) \le \frac{2}{t}$.

Intuitively, the support of an (r,t)-regular distribution can be partitioned into t sets, each of which has the same radius bound and comparable mass. This notion can be viewed as a distributional counterpart of a regular instance in [29]. The next lemma is our main technical lemma, which establishes a coreset construction for regular distributions.

Lemma 6. *Assume p is an (r,t)-regular distribution such that $t \le k$, then with probability $1 - \mathrm{poly}(\frac{\varepsilon}{k \log n})$, Algorithm 1 computes an $(\varepsilon, k, \varepsilon r)$-coreset of size $\tilde{O}(k\varepsilon^{-4} \log n)$ of p.*

Theorem 3, which concerns general distributions, follows from Lemma 6 via a reduction based on ring decompositions. This argument extends the ring decomposition framework in [15].

Proof of Theorem 3. We view p as a dataset on (M, ρ) and assume we are given an $O(1)$-approximation $Q \subseteq P := \mathrm{supp}(p)$ for k-MEDIANS on p. The computation time needed to obtain Q depends on the application. Assign each point of P to a nearest center in Q, breaking ties arbitrarily. Let $C(y) \subseteq P$ denote the set of points assigned to y and let q be a distribution supported on Q defined as $q(y) = p(C(y))$ for each $y \in Q$. As Q is an $O(1)$-approximation for k-MEDIANS, we know that for every distribution w such that $|\mathrm{supp}(w)| \le k$, we have $\mathrm{EMD}(p,q) \le O(1) \cdot \mathrm{EMD}(p,w)$. Let $\delta = \mathrm{EMD}(p,q)$.

Our coreset construction of $\widehat{p}$ proceeds in three stages.

In the first stage, we set $g = \max\{\varepsilon\delta, \rho(P \setminus Q, Q)/2\}$. For every $y \in Q$ such that $p(B(y,g) \cap C(y)) > 0$, we include y in $\mathrm{supp}(\widehat{p})$ with $\widehat{p}(y) = p(B(y,g) \cap C(y))$.

In the second stage, we decompose P into $\ell = \lceil \log(2\tau/\varepsilon) \rceil$ sets $A_1, \dots, A_\ell$ by setting $A_i = \bigcup_{y \in Q}(C(y) \cap (B(y, 2^i g) \setminus B(y, 2^{i-1} g)))$, $i \in [\ell]$.

We now show that $A_1, \dots, A_\ell$ cover all the remaining points in P. If $\tau = 1/p_{\min}$, since $p_{\min} \max_{x \in P} \rho(x,Q) \le \mathrm{EMD}(p,q) = \delta$, we have $\max_{x \in P} \rho(x,Q) \le \delta\tau \le 2^\ell \varepsilon\delta$, which implies that $\bigcup_{i \in [\ell]} A_i$ covers the remaining points in P. If $\tau = \lambda$, since $Q \subseteq P$ and $g \ge \rho(P \setminus Q, Q)/2$, we know that $\max_{x \in P} \rho(x,Q) \le 2\tau g \le 2^\ell g$ for $0 < \varepsilon \le 1$, hence, $\bigcup_{i \in [\ell]} A_i$ covers the remaining points in P.

In the third stage, we decompose each $p|_{A_i}$ ($i \in [\ell]$) into multiple regular distributions. Let q_i be a measure supported on Q defined by $q_i(y) = p(C(y) \cap (B(y, 2^i g) \setminus B(y, 2^{i-1} g)))$. Let $\delta_i = \mathrm{EMD}(p|_{A_i}, q_i)$. Then $\delta_i \ge 2^{i-1} g \cdot p(A_i)$ and $\sum_{i=1}^{\ell} \delta_i \le \delta$. We further decompose A_i into $h+1$ sets $A_{i,0}, \dots, A_{i,h}$, where $h = \lceil \log(k/\varepsilon) \rceil$, such that

$$A_{i,0} = \bigcup_{y \in Q : q_i(y) \in (0, \varepsilon p(A_i)/k]} \big(C(y) \cap (B(y, 2^i g) \setminus B(y, 2^{i-1} g))\big),$$

and

$$A_{i,j} = \bigcup_{y \in Q : q_i(y) \in (2^{j-1}\varepsilon p(A_i)/k, 2^j \varepsilon p(A_i)/k]} \big(C(y) \cap (B(y, 2^i g) \setminus B(y, 2^{i-1} g))\big).$$

For each $y \in Q$ with $q_i(y) \in (0, \varepsilon p(A_i)/k]$, we add $p(C(y) \cap (B(y, 2^i g) \setminus B(y, 2^{i-1} g)))$ to $\widehat{p}(y)$. For each $j \in [h]$ with $p(A_{i,j}) > 0$, we normalize $p|_{A_{i,j}}$ by dividing by $p(A_{i,j})$ to obtain a probability distribution $\tilde{p}_{i,j}$. This distribution is $(2^i g, t_{i,j})$-regular, where $t_{i,j} \le k$ is the number of centers $y \in Q$ contributing to $A_{i,j}$: each part lies in $B(y, 2^i g)$, and the masses of the parts are within a factor of 2 of each other. We apply the algorithm of Lemma 6 to construct an $(\varepsilon, k, \varepsilon \cdot 2^i g)$ coreset $\mu_{i,j}$ for $\tilde{p}_{i,j}$, then add the scaled measure $p(A_{i,j})\mu_{i,j}$ to $\widehat{p}$.

The coreset property of $\widehat{p}$ follows from the mergeability of coresets and a union bound. The size of the coreset can be bounded by

$$|\mathrm{supp}(\widehat{p})| \le \underbrace{k}_{\text{from Stage 1}} + \underbrace{k\ell}_{\text{from } A_{i,0}} + \underbrace{\ell h \tilde{O}(\varepsilon^{-4} k \log n)}_{\text{from } A_{i,j} \text{ and Lemma 6}} = \tilde{O}\left(\varepsilon^{-4} k \log n \log \frac{\tau}{\varepsilon}\right).$$

The total additive error can be bounded by (note that the error in the first stage is zero when $\rho(P \setminus Q, Q)/2 > \varepsilon\delta$)

$$\underbrace{\varepsilon\delta}_{\text{from Stage 1}} + \sum_{i=1}^{\ell} \underbrace{k \cdot 2^i g \cdot \frac{\varepsilon p(A_i)}{k}}_{\text{from } A_{i,0}} + \sum_{i=1}^{\ell} \sum_{j=1}^{h} \underbrace{p(A_{i,j}) \cdot \varepsilon 2^i g}_{\text{from } A_{i,j}} \le \varepsilon\delta + 2 \sum_{i=1}^{\ell} \varepsilon 2^i g \cdot p(A_i)$$

$$\le \varepsilon\delta + 4 \sum_{i=1}^{\ell} \varepsilon\delta_i = O(\varepsilon\delta),$$

which can be transformed into a multiplicative error after rescaling ε. □

Algorithm 1 $\mathrm{Coreset}(p, r, k, \varepsilon, \{a_1, \cdots, a_t\})$

Require: An (r, t)-regular distribution p with parts $B_i \subseteq B(a_i, r)$, $i \in [t]$.
Ensure: Coreset μ
1: $m \leftarrow \tilde{O}(k\varepsilon^{-4} \log n)$
2: $U \leftarrow$ an i.i.d. sample of size m from p
3: **for** $i = 1, \ldots, t$ **do**
4: $\quad u_i \leftarrow |U \cap B_i|$
5: $\quad$ for each $x \in U \cap B_i$, set $\mu(x) = \frac{p(B_i)}{u_i}$
6: **end for**
7: **return** μ

5 Proof of Lemma 6

Our proof strategy consists of two steps. First, we show for a fixed k-sparse distribution q that $\mathrm{EMD}(\mu, q) \approx \mathrm{EMD}(p, q)$ with high probability. Second, we employ a net argument over a net of k-sparse distributions to prove the coreset property, i.e. the EMD approximation holds for every k-sparse distribution.

In the remainder of this section, we fix a distribution q with $|\operatorname{supp}(q)| = k$. Let $P = \operatorname{supp}(p)$, $Q = \operatorname{supp}(q)$ and $\Lambda : P \times Q \to \mathbb{R}_{\geq 0}$ denote the optimal transportation matrix from p to q. Recall that p is an (r, t)-regular distribution with parts $B_1 \subseteq B(a_1, r), \ldots, B_t \subseteq B(a_t, r)$. For each $x \in P$, let $c(x)$ denote the center a_i of the unique part B_i containing x.

Our coreset construction algorithm is Algorithm 1, which samples $\tilde{O}(k\varepsilon^{-4} \log n)$ points and scales the weights (Line 5) to preserve the total mass of each B_i. We remark that such a scaling step is widely employed in constructing coresets for constrained clustering [8, 12, 30, 29].

The following result gives the basic property of the output distribution μ. It is an adaptation of Lemma 19 of [19]; see also Lemma 6.8 of [29].

Lemma 7 ([19, 29])**.** *It holds that $\mu(B_i) = p(B_i)$ for every $i \in [t]$. Furthermore, with probability at least $1 - \mathrm{poly}(\frac{\varepsilon}{k \log n})$, it holds that $\mu(x) \in [\frac{1-\varepsilon}{m}, \frac{1+\varepsilon}{m}]$ for all $x \in \operatorname{supp}(\mu)$.*

Close and Far Transportation To analyse the transportation cost, we decompose the optimal transportation Λ into close and far parts based on distance.

We define the *close* transportation matrix Λ_c by

$$\Lambda_c(x, y) := \begin{cases} \Lambda(x, y) & \rho(c(x), y) \leq \varepsilon^{-1} r \\ 0 & \text{otherwise} \end{cases}, \quad \forall x \in P, y \in Q,$$

and let the *far* transportation matrix be $\Lambda_f := \Lambda - \Lambda_c$.

Based on this decomposition, we define measures p_c and p_f on P by

$$p_c(x) := \sum_{y \in Q} \Lambda_c(x, y), \qquad p_f(x) := p(x) - p_c(x), \quad \forall x \in P,$$

and similarly define measures q_c and q_f on Q by

$$q_c(y) := \sum_{x \in P} \Lambda_c(x, y), \qquad q_f(y) := q(y) - q_c(y), \quad \forall y \in Q.$$

Next, we establish bounds on three important quantities: the close transportation cost under μ compared with that under p, and two L_1-distances that measure the difference between μ and p on the close and far parts, respectively.

Close Transportation Cost For every $x \in P$, we define $v(x) = \sum_{y \in Q} \frac{\Lambda_c(x,y)\rho(x,y)}{p(x)}$ to be the connection cost from x to its close neighbors in the optimal assignment. We have the following estimate; its proof is deferred to Section B.

Lemma 8. *With probability at least* $1 - n^{-\Omega(k \log(k/\varepsilon))}$*, it holds that*

$$\sum_{x \in \mathrm{supp}(\mu)} \mu(x) v(x) \in (1 \pm \varepsilon) \sum_{x \in P} p(x) v(x) \pm \varepsilon r.$$

L_1-distances Define

$$L_1^c = \sum_{y \in Q} \left| q_c(y) - \sum_{x \in \mathrm{supp}(\mu)} \frac{\mu(x) \Lambda_c(x, y)}{p(x)} \right| \quad \text{and} \quad L_1^f = \sum_{i \in [t]} \left| p_f(B_i) - \sum_{x \in \mathrm{supp}(\mu) \cap B_i} \frac{\mu(x) p_f(x)}{p(x)} \right|.$$

Lemma 9. *With probability at least* $1 - n^{-\Omega(k \log(k/\varepsilon))}$*, it holds that* $L_1^c \le \varepsilon^2$ *and* $L_1^f \le \varepsilon^2$*.*

The proof of Theorem 9 is deferred to Section C.

5.1 Comparing $\mathrm{EMD}(\mu, q)$ and $\mathrm{EMD}(p, q)$

With the bounds on close transportation cost and L_1-distances established, we now proceed to compare $\mathrm{EMD}(\mu, q)$ and $\mathrm{EMD}(p, q)$. This comparison uses standard results from optimal transport, which we first recall. Lemmas 10 and 11 are classical and their proofs can be found, for example, in [43].

Definition 5.1 (Cyclical Monotonicity)**.** Let A be a transportation matrix from p to q. The matrix A is said to be *cyclically monotone* if for every $t \ge 1$ and every sequence $(x_1, y_1), \ldots, (x_t, y_t) \in \mathrm{supp}(A)$, it holds that $\sum_{i=1}^{t} \rho(x_i, y_i) \le \sum_{i=1}^{t} \rho(x_i, y_{i+1})$, where $y_{t+1} = y_1$.

For transport matrices between distributions with finite supports, cyclical monotonicity is equivalent to optimality.

Lemma 10. *A transportation matrix A is optimal if and only if it is cyclically monotone.*

Moreover, if A is cyclically monotone, then any transportation matrix B with $\mathrm{supp}(B) \subseteq \mathrm{supp}(A)$ is also cyclically monotone. We therefore have the following lemma.

Lemma 11. *Let w be a subdistribution of p and $\Pi : \mathrm{supp}(w) \times \mathrm{supp}(q) \to \mathbb{R}_{\ge 0}$ be a transportation matrix from w to q. If $\mathrm{supp}(\Pi) \subseteq \mathrm{supp}(\Lambda)$, then Π is an optimal transportation matrix from w to q.*

We now use these classical facts to relate $\mathrm{EMD}(\mu, q)$ to $\mathrm{EMD}(p, q)$. It is natural to define a (semi-)transportation matrix Π from μ to q by setting $\Pi(x, y) = \mu(x)\Lambda(x, y)/p(x)$ for all $x \in \mathrm{supp}(\mu)$ and $y \in Q$. However, since μ is random, we do not necessarily have $\sum_{x \in \mathrm{supp}(\mu)} \Pi(x, y) = q(y)$ for every $y \in Q$, that is, Π may not be a true transportation matrix. Nevertheless, we show that μ can be modified, at an affordable error, to another subdistribution w, and Λ can be correspondingly adjusted to a transportation matrix Θ from w to q that satisfies the condition of Lemma 11.

The adjustment is guided by the following observation: for the far transportations, it suffices to preserve the *total* far transportation mass leaving each regularity ball, rather than the mass leaving each individual point. Indeed, redistributing far transportation mass among points of the same ball B_i moves each unit of mass by at most $2r$, whereas every far transportation from B_i travels a distance of at least $\varepsilon^{-1} r - r$. Hence such a redistribution changes the far transportation cost by at most an $O(\varepsilon)$ fraction. The close transportation cost and the discrepancies in the marginals are controlled separately by the concentration bounds in Lemmas 8 and 9. This is why the adjustments below keep the total far mass leaving each ball unchanged, while allowing small changes to the close transportations.

Lemma 12. *With probability at least* $1 - n^{-\Omega(k \log(k/\varepsilon))}$*, there exist a subdistribution w of p and a transportation matrix $\Theta : \mathrm{supp}(w) \times Q \to \mathbb{R}_{\ge 0}$ such that the following hold.*

(i) $\mathrm{EMD}(w, q) = \sum_{x \in \mathrm{supp}(w), y \in Q} \Theta(x, y) \rho(x, y)$*;*

(ii) $(1 - C_1\varepsilon)\,\mathrm{EMD}(w, q) - C_2 \varepsilon r \le \mathrm{EMD}(\mu, q) \le (1 + C_1\varepsilon)\,\mathrm{EMD}(w, q) + C_2 \varepsilon r$*;*

(iii) $(1 - C_3\varepsilon)\,\mathrm{EMD}(p, q) - C_4 \varepsilon r \le \mathrm{EMD}(w, q) \le (1 + C_3\varepsilon)\,\mathrm{EMD}(p, q) + C_4 \varepsilon r$*.*

Here, $C_1, C_2, C_3, C_4 > 0$ are absolute constants.

Proof. We condition on the good events of Theorems 8 and 9, which happen with probability at least $1 - n^{-\Omega(k \log(k/\varepsilon))}$.

Initially, we set $w = \mu$ and $\Theta(x, y) = \mu(x)\Lambda(x, y)/p(x)$ for all $x \in \mathrm{supp}(w)$ and $y \in Q$, and shall modify w in three stages. Let

$$\Theta_c(x, y) := \begin{cases} \Theta(x, y) & \rho(c(x), y) \le \varepsilon^{-1} r \\ 0 & \text{otherwise} \end{cases}$$

and $\Theta_f(x, y) = \Theta(x, y) - \Theta_c(x, y)$.

In the first stage, we redistribute mass (between close and far transportations) within each part B_i ($i \in [t]$) so that

$$p_f(B_i) = \sum_{x \in B_i \cap \mathrm{supp}(w)} \sum_{y \in Q} \Theta_f(x, y), \quad i \in [t].$$

By Lemma 9 and the fact that $\mathrm{diam}(B_i) \le 2r$, this modification can be achieved by moving at most ε^2 total mass over a distance at most $2r$. The incurred error in earth mover's distance is at most $2\varepsilon^2 r$.

In the second stage, our goal is to ensure that $q_c(y) = \sum_{x \in \mathrm{supp}(w)} \Theta_c(x, y)$ for all $y \in Q$ by adjusting only within close transportations (adjustments across balls are allowed). From Lemma 9, together with the fact at most ε^2 total mass was adjusted in first stage, we have

$$\sum_{y \in Q} \left| q_c(y) - \sum_{x \in \mathrm{supp}(w)} \Theta_c(x, y) \right| \le 2\varepsilon^2.$$

Thus, correcting $q_c(\cdot)$ will adjust at most $2\varepsilon^2$ total mass. Since $\Theta_c(x, y)$ corresponds to close connections, the mass is moved over a distance at most $2r/\varepsilon$. Consequently, the total error incurred in this stage can be bounded by $2\varepsilon^{-1} r \cdot 2\varepsilon^2 = O(\varepsilon r)$.

In the third stage, we ensure $q_f(y) = \sum_{x \in \mathrm{supp}(w)} \Theta_f(x, y)$ for all $y \in Q$ by adjusting only within far transportations (adjustments across balls are allowed). We bound the error in this stage later.

Throughout all three stages we can preserve $\mathrm{supp}(\Theta) \subseteq \mathrm{supp}(\Lambda)$, so by Lemma 11, we have $\mathrm{EMD}(w, q) = \sum_{x \in \mathrm{supp}(w)} \sum_{y \in Q} \Theta(x, y)\rho(x, y)$, which establishes Item (i).

Next we bound the error incurred in the third stage. Recall that $\Theta_f(x, y) > 0$ only if $\rho(c(x), y) > \varepsilon^{-1} r$. Thus, all points within the same part B_i can be treated as effectively the same point at the cost of an $O(\varepsilon)$ multiplicative distortion in distance. Consequently, the third stage contributes a total error of at most $O(\varepsilon) \cdot \mathrm{EMD}(w, q)$. Combining with the error of $O(\varepsilon r)$ from the first two stages, we establish Item (ii).

Last, we prove Item (iii). The error in the far transportations can be bounded as in the third stage, leading to $O(\varepsilon) \cdot \mathrm{EMD}(p, q)$. The error in the close transportations follows from Lemma 8. Combining the two bounds completes the proof. □

The desired EMD approximation for a fixed q now follows directly from the triangle inequality and Lemma 12:

$$\begin{aligned} |\mathrm{EMD}(\mu, q) - \mathrm{EMD}(p, q)| &\le |\mathrm{EMD}(\mu, q) - \mathrm{EMD}(w, q)| + |\mathrm{EMD}(w, q) - \mathrm{EMD}(p, q)| \\ &= (O(\varepsilon)\,\mathrm{EMD}(w, q) + O(\varepsilon r)) + (O(\varepsilon)\,\mathrm{EMD}(p, q) + O(\varepsilon r)) \\ &= O(\varepsilon)\,\mathrm{EMD}(p, q) + O(\varepsilon r). \end{aligned}$$

5.2 Net Argument

In the previous subsection we established, after rescaling k and ε, that

$$\mathrm{EMD}(\mu, q) \in (1 \pm \varepsilon)\,\mathrm{EMD}(p, q) \pm \varepsilon r$$

with high probability for any fixed $(2k)$-sparse distribution. Now we handle the second step of the proof: extending this approximation guarantee from a fixed $(2k)$-sparse distribution to all k-sparse distributions via a net-style argument. For each k-sparse distribution, we reduce the EMD

approximation to an auxiliary $(2k)$-sparse distribution, and we cover all such auxiliary distributions by a discretization family that plays the role of a net. For convenience, we shall still refer to this discretization family as "a net".

We need set partitions in our net construction.

Definition 5.2 (Set Partitions)**.** Let N be a positive integer. A set partition of $[N]$ is a collection of disjoint nonempty subsets $S_1, \ldots, S_\ell \subseteq [N]$ such that $S_1 \cup \cdots \cup S_\ell = [N]$. Let $\mathbf{P}_N$ be the set of all partitions of $[N]$.

Fact 13. *$|\mathbf{P}_N| = B_N$, where B_N is the N-th Bell number and $B_N < N^N$.*

Construction of Net Let $l = \lfloor k^2\varepsilon^{-2} \rfloor$, $\alpha = 1/l$ and $N = t + k$. Let $\mathcal{N}$ denote a set of N-sparse distributions, where each distribution in $\mathcal{N}$ is described by the following tuple

$$\theta = (\mathcal{S}, b_1, \ldots, b_k, n_1, \ldots, n_N).$$

Here, $b_1, \ldots, b_k \in M$, $\mathcal{S} \in \mathbf{P}_N$ describes a partition of $\{b_1, \ldots, b_k, a_1, \ldots, a_t\}$ and the numbers $n_1, \ldots, n_N \in \mathbb{R}_{\geq 0}$ satisfy that for every $S \in \mathcal{S}$,

$$\sum_{i \in S \cap [t]} p(B_i) = \sum_{i \in S} n_i,$$

and for each $i \in S$, n_i is a multiple of $\alpha \sum_{j \in S \cap [t]} p(B_j)$. The size of $\mathcal{N}$ is bounded by

$$|\mathcal{N}| \leq (2k)^{2k} \cdot n^k \cdot (k^2\varepsilon^{-2} + 1)^{2k} \leq \exp\big(O(k \log(nk/\varepsilon))\big).$$

Intuitively, θ describes a distribution q supported on $(k + t)$ points $\{b_1, \ldots, b_k, a_1, \ldots, a_t\}$ such that $q(b_i) = n_i$ for each $i \in [k]$ and $q(a_i) = n_{i+k}$ for each $i \in [t]$. Moreover, $\mathcal{S}$ describes a partition of these $t + k$ points in $\mathrm{supp}(q)$.

Net Argument The last piece of the proof is to show that if the EMD approximation holds for all distributions in the net $\mathcal{N}$, it must hold for all k-sparse distributions q. This is formalized as the following lemma, whose proof is deferred to Section D.

Lemma 14. *Let μ be a sub-distribution of p. If $\mathrm{EMD}(\mu, q) \in (1 \pm \varepsilon) \cdot \mathrm{EMD}(p, q) \pm \varepsilon r$ for every $q \in \mathcal{N}$, then μ is an $(C_1\varepsilon, k, C_2\varepsilon r)$-coreset of p for some absolute constants $C_1, C_2 > 0$.*

Acknowledgment

C.W. In is supported by a NTU Research Scholarship. Y. Li is supported in part by Singapore Ministry of Education AcRF Tier 1 grant RG21/25. X. Wu is supported in part b

References

[1] Pankaj K. Agarwal, Sariel Har-Peled, and Kasturi R. Varadarajan. Approximating extent measures of points. *J. ACM*, 51(4):606–635, July 2004.

[2] Pankaj K. Agarwal, Sharath Raghvendra, Pouyan Shirzadian, and Keegan Yao. A combinatorial algorithm for the semi-discrete optimal transport problem. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, 2024.

[3] Pankaj K. Agarwal, Sharath Raghvendra, Pouyan Shirzadian, and Keegan Yao. Fast and accurate approximations of the optimal transport in semi-discrete and discrete settings. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 4514–4529, 2024.

[4] Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. Earth mover distance over high-dimensional spaces. In *SODA*, volume 8, pages 343–352, 2008.

[5] Alexandr Andoni and Hengjie Zhang. Sub-quadratic (1+ε)-approximate Euclidean spanners, with applications. In *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 98–112, 2023.

[6] Franz Aurenhammer, Friedrich Hoffmann, and Boris Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998.

[7] Daniel Baker, Vladimir Braverman, Lingxiao Huang, Shaofeng H-C Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in graphs of bounded treewidth. In *International Conference on Machine Learning*, pages 569–579. PMLR, 2020.

[8] Sayan Bandyapadhyay, Fedor V. Fomin, and Kirill Simonov. On coresets for fair clustering in metric and Euclidean spaces and their applications. In *ICALP*, volume 198 of *LIPIcs*, pages 23:1–23:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

[9] Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. Fair algorithms for clustering. *Advances in Neural Information Processing Systems*, 32, 2019.

[10] Lorenzo Beretta, Vincent Cohen-Addad, Rajesh Jayaram, and Erik Waingarten. Approximating high-dimensional earth mover's distance as fast as closest pair. *FOCS*, 2025.

[11] Lorenzo Beretta and Aviad Rubinstein. Approximate earth mover's distance in truly-subquadratic time. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 47–58, 2024.

[12] Vladimir Braverman, Vincent Cohen-Addad, H-C Shaofeng Jiang, Robert Krauthgamer, Chris Schwiegelshohn, Mads Bech Toftrup, and Xuan Wu. The power of uniform sampling for coresets. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 462–473. IEEE, 2022.

[13] Vladimir Braverman, Shaofeng H.-C. Jiang, Robert Krauthgamer, and Xuan Wu. Coresets for clustering in excluded-minor graphs and beyond. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2679–2696. SIAM, 2021.

[14] Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thiry-Fourth annual ACM Symposium on Theory of Computing*, pages 380–388, 2002.

[15] Ke Chen. On coresets for k-median and k-means clustering in metric and Euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947, 2009.

[16] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[17] Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, Chris Schwiegelshohn, and Omar Ali Sheikh-Omar. Improved coresets for Euclidean k-means. *Advances in Neural Information Processing Systems*, 35:2679–2694, 2022.

[18] Vincent Cohen-Addad, Shaofeng H.-C. Jiang, Qiaoyuan Yang, Yubo Zhang, and Samson Zhou. Fair clustering in the sliding window model. In *Proceedings of ICLR 2025*, 2025.

[19] Vincent Cohen-Addad, Kasper Green Larsen, David Saulpic, and Chris Schwiegelshohn. Towards optimal lower bounds for k-median and k-means coresets. In *Proceedings of STOC*, pages 1038–1051. ACM, 2022.

[20] Vincent Cohen-Addad, David Saulpic, and Chris Schwiegelshohn. A new coreset framework for clustering. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 169–182, 2021.

[21] Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 569–578, 2011.

[22] Dan Feldman and Leonard J Schulman. Data reduction for weighted and outlier-resistant clustering. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1343–1354. SIAM, 2012.

[23] Sariel Har-Peled, Piotr Indyk, and Anastasios Sidiropoulos. Euclidean spanners in high dimensions. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 804–809. SIAM, 2013.

[24] Sariel Har-Peled and Akash Kushal. Smaller coresets for k-median and k-means clustering. *Discret. Comput. Geom.*, 37(1):3–19, 2007.

[25] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of STOC*, pages 291–300. ACM, 2004. `https://arxiv.org/abs/1810.12826`.

[26] Valentin Hartmann and Dominic Schuhmacher. Semi-discrete optimal transport: a solution procedure for the unsquared euclidean distance case. *Mathematical Methods of Operations Research*, 92(1):133–163, 2020.

[27] Lingxiao Huang, Shaofeng H.-C. Jiang, Jianing Lou, and Xuan Wu. Near-optimal coresets for robust clustering. In *Proceedings of ICLR 2023*, 2023.

[28] Lingxiao Huang, Shaofeng H.-C. Jiang, and Nisheeth K. Vishnoi. Coresets for clustering with fairness constraints. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

[29] Lingxiao Huang, Zhenyu Jiang, Yi Li, and Xuan Wu. Towards tight robust coresets for k-medians clustering. *ICALP*, 2026.

[30] Lingxiao Huang, Jian Li, Pinyan Lu, and Xuan Wu. Coresets for constrained clustering: General assignment constraints and improved size bounds. In *Proceedings of the 2025 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 4732–4782. SIAM, 2025.

[31] Lingxiao Huang, Jian Li, and Xuan Wu. On optimal coreset construction for euclidean (k, z)-clustering. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1594–1604, 2024.

[32] Lingxiao Huang and Nisheeth K. Vishnoi. Coresets for clustering in Euclidean spaces: importance sampling is nearly optimal. In *Proceedings of STOC*, pages 1416–1429. ACM, 2020.

[33] Chansophea Wathanak In, Yi Li, David Woodruff, and Xuan Wu. Robust sparsification via sensitivity. In *International Conference on Machine Learning*, pages 26446–26463. PMLR, 2025.

[34] Piotr Indyk and Thaper Nitin. Fast image retrieval via embeddings. *Workshop on Statistical and Computational Theories of Vision (at ICCV)*, 2003.

[35] Shaofeng H.-C. Jiang and Jianing Lou. Coresets for robust clustering via black-box reductions to vanilla case. In *ICALP*, volume 334 of *LIPIcs*, pages 101:1–101:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2025.

[36] Andrey Boris Khesin, Aleksandar Nikolov, and Dmitry Paramonov. Preconditioning for the geometric transportation problem. *Journal of Computational Geometry*, 11(2):234–259, Mar. 2021.

[37] Max Dupré la Tour and David Saulpic. Almost-linear time approximation algorithm to euclidean k-median and k-means. *arXiv preprint arXiv:2407.11217*, 2024.

[38] Michael Langberg and Leonard J. Schulman. Universal epsilon-approximators for integrals. In *SODA*, pages 598–607. SIAM, 2010.

[39] Quentin Mérigot. A multiscale approach to optimal transport. In *Computer graphics forum*, volume 30, pages 1583–1592. Wiley Online Library, 2011.

[40] Shyam Narayanan and Jelani Nelson. Optimal terminal dimensionality reduction in euclidean space. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 1064–1069, 2019.

[41] Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. Fair coresets and streaming algorithms for fair k-means. In *International Workshop on Approximation and Online Algorithms*, pages 232–251. Springer, 2019.

[42] Pravin Vaidya. Geometry helps in matching. In *Proceedings of the twentieth annual ACM symposium on theory of computing*, pages 422–425, 1988.

[43] Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

[44] Arman Yousefi and Rafail Ostrovsky. Improved approximation algorithms for earth-mover distance in data streams. *arXiv preprint arXiv:1404.6287*, 2014.

A Missing proofs in Section 1.2

Proof of Corollary 4. We note that Lemma 6 still holds for k-sparse distributions in $\mathbb{R}^d$. Hence, the only difference lies in the construction of the ε-net.

Let N_ε denote an (εr)-net of $B(a_1, \varepsilon^{-1}r) \cup \cdots \cup B(a_t, \varepsilon^{-1}r)$, with respect to the Euclidean distance, then $|N_\varepsilon| \le k\varepsilon^{-O(d)}$. We then construct an ε-net $\mathcal{N}$ for EMD by following the same construction in Section 5.2, in which the underlying metric space M is replaced with the net N_ε.

It remains to adapt the proof of Lemma 14. The difference in Euclidean spaces is that the set of possible center locations can be infinite. Nevertheless, we can still construct an auxiliary distribution $\hat{q}$ supported on $\{a_1, \ldots, a_t\} \cup N_\varepsilon$ and show it suffices to prove $\mathrm{EMD}(\mu, q) \le (1+O(\varepsilon))\,\mathrm{EMD}(\mu, \hat{q}) + O(\varepsilon r)$. The rest of the proof is identical.

To bound the coreset size, we observe that $\log |N_\varepsilon| \le O(d \log(k/\varepsilon))$ and we can obtain a coreset size of $\tilde{O}(kd\varepsilon^{-4} \log n)$ through the reduction in Section 4. We can further use the terminal embedding [40] and iterative size reduction [13] to replace d by an $\varepsilon^{-2} \log^{O(1)}(k/\varepsilon)$ factor[5], which is standard in the literature. The resulting coreset size is $\tilde{O}(k\varepsilon^{-6} \log n)$.

For the running time, we note that the bottleneck is the computation of an approximate solution for k-MEDIANS. We use a recent result of [37], which gives a $\mathrm{polylog}(k)$-approximation in $\tilde{O}(nd)$ time. We can scale ε with the $\mathrm{polylog}(k)$ factor and still obtain a coreset of size $\tilde{O}(k\varepsilon^{-6} \log n)$. □

Theorem 5 is a direct application of Corollary 4. We employ a classical reduction (see Theorem 4.3 of [28]), which reduces the construction of fair coresets for multiple groups to the case of a single group; the latter is equivalent to our definition of a coreset. Based on such coreset construction, the streaming algorithm can be obtained via a standard merge-and-reduce framework [25, 15].

B Proof of Theorem 8

Proof. Since U consists of m i.i.d. samples from p, we have $\mathbb{E}[\frac{1}{m}\sum_{x\in U} v(x)] = \sum_{x\in P} p(x)v(x)$. By Lemma 7, we have $\sum_{x\in \mathrm{supp}(\mu)} \mu(x)v(x) = \sum_{x\in \mathrm{supp}(\mu)} \frac{1}{m} \cdot \frac{mp(x)}{u_i} \cdot v(x) \in (1 \pm \varepsilon^2) \cdot \frac{1}{m}\sum_{x\in U} v(x)$. Thus,

$$\mathbb{E}\left[\sum_{x\in \mathrm{supp}(\mu)} \mu(x)v(x)\right] \in (1 \pm \varepsilon^2) \sum_{x\in P} p(x)v(x).$$

[5] Specifically, for a data set $X \subset \mathbb{R}^d$, one can apply the terminal embedding of [40] with a target dimension $d' = \tilde{O}(\varepsilon^{-2} \log |X|)$ and construct a coreset of X in this lower-dimensional space. Using Theorem 3.1 of [13], one can further reduce the $\log |X|$ factor by iteratively constructing a coreset of a coreset with different values of ε. We remark that this technique cannot remove the $\log n$ factor in the coreset size, since here n denotes the total weight (of the original unweighted dataset), which can be much larger than the cardinality $|X|$ when X is weighted.

Next we control the deviation, for which we employ Hoeffding's inequality. Note that each $v(x) \in O(\varepsilon^{-1}r)$ and $\mu(x) \in \Theta(\frac{1}{m})$, we know that

$$\begin{aligned}\mathbb{P}\left[\left|\sum_{x\in\operatorname{supp}(\mu)}\mu(x)v(x) - \mathbb{E}\sum_{x\in\operatorname{supp}(\mu)}\mu(x)v(x)\right| > \varepsilon r\right] &\le \exp\left\{-\Omega\left(\frac{\varepsilon^2 r^2}{m\cdot(\frac{1}{m}\varepsilon^{-1}r)^2}\right)\right\}\\ &= \exp\{-\Omega(m\varepsilon^4)\}\\ &\le \frac{1}{n^{\Omega(k\log(k/\varepsilon))}},\end{aligned}$$

provided $m \gtrsim k\varepsilon^{-4}\log n\log(k/\varepsilon)$. The result follows by a rescaling of ε. □

C Proof of Theorem 9

Proof. For a subset $I \subseteq Q$, define a connection vector $v_I : P \to \mathbb{R}_{\ge 0}$ as $v_I(x) = \sum_{y\in I}\Lambda_c(x,y)$. By separating positive and negative discrepancies, we obtain

$$L_1^c \le 2\max_{I\subseteq Q}\left|\sum_{x\in\operatorname{supp}(\mu)}\frac{\mu(x)v_I(x)}{p(x)} - \sum_{x\in P}v_I(x)\right|.$$

Since U consists of m i.i.d. samples from p, we know that

$$\mathbb{E}\left[\frac{1}{m}\sum_{x\in U}\frac{v_I(x)}{p(x)}\right] = \sum_{x\in P}v_I(x).$$

It follows from Lemma 7 that

$$\mathbb{E}\left[\sum_{x\in\operatorname{supp}(\mu)}\frac{\mu(x)v_I(x)}{p(x)}\right] \in (1\pm\varepsilon^2)\sum_{x\in P}v_I(x).$$

Since $\sum_{x\in P}v_I(x) \le 1$, we have

$$\mathbb{E}\left[\sum_{x\in\operatorname{supp}(\mu)}\frac{\mu(x)v_I(x)}{p(x)}\right] \in \sum_{x\in P}v_I(x) \pm \varepsilon^2.$$

It remains to control the deviation and scale ε, for which we employ Hoeffding's inequality. Note that $\mu(x) \in \Theta(\frac{1}{m})$ and $v_I(x)/p(x) \in [0,1]$, so for a fixed $I \subseteq Q$, we have

$$\begin{aligned}\mathbb{P}\left[\left|\sum_{x\in\operatorname{supp}(\mu)}\frac{\mu(x)v_I(x)}{p(x)} - \mathbb{E}\sum_{x\in\operatorname{supp}(\mu)}\frac{\mu(x)v_I(x)}{p(x)}\right| > \varepsilon^2\right] &\le \exp\left\{-\Omega\left(\frac{\varepsilon^4}{m\cdot(\frac{1}{m})^2}\right)\right\}\\ &= \exp\{-\Omega(m\varepsilon^4)\}\\ &\le n^{-\Omega(k\log(k/\varepsilon))},\end{aligned}$$

provided $m \gtrsim k\varepsilon^{-4}\log n\log(k/\varepsilon)$.

Since there are 2^k possible subsets I, a union bound shows that the bound above holds simultaneously for all $I \subseteq Q$, still with probability at least $1 - n^{-\Omega(k\log(k/\varepsilon))}$.

An analogous argument also shows that $L_1^f \le \varepsilon^2$. □

D Proof of Lemma 14

Proof. Fix a k-sparse distribution q. We shall show that there exists a transportation matrix Π from μ to q such that

$$\sum_{x\in\operatorname{supp}(\mu)}\sum_{y\in\operatorname{supp}(q)}\Pi(x,y)\cdot\rho(x,y) \le (1+O(\varepsilon))\operatorname{EMD}(p,q) + O(\varepsilon r),$$

and that there exists a transportation matrix Π' from p to q such that

$$\sum_{x\in\mathrm{supp}(p)}\sum_{y\in\mathrm{supp}(q)}\Pi'(x,y)\cdot\rho(x,y)\le(1+O(\varepsilon))\,\mathrm{EMD}(\mu,q)+O(\varepsilon r).$$

Since the two inequalities are proved in exactly the same way, we present only the proof of the first.

Let Λ denote the optimal transportation matrix from p to q, and let Λ_c, p_c, q_c be as defined in Section 5. Let $A=\{a_1,\ldots,a_t\}$ be the set of centers, and define a measure q_a on A as $q_a(a_i)=p_f(B_i)$ for all $i\in[t]$. Finally, we define a $(k+t)$-sparse distribution $\widehat{q}=q_c+q_a$.

We first observe that $\mathrm{EMD}(p,\widehat{q})+\mathrm{EMD}(q_a,q_f)\le(1+O(\varepsilon))\,\mathrm{EMD}(p,q)$. To see this, we slightly modify the transportation process in the far part. Whenever mass is transported from a point $x\in\mathrm{supp}(p)$ to a point $y\in\mathrm{supp}(q)$ with $\rho(c(x),y)>\varepsilon^{-1}r$, we perform two steps: first transport the mass from x to the center $c(x)$, and then from $c(x)$ to y. Since $\rho(x,c(x))\le r\le\varepsilon\cdot\rho(c(x),y)$, this rerouting only increases the total cost by an $O(\varepsilon)$-fraction. Keeping the close transportations unchanged and applying the first step to the far transportations gives a feasible plan from p to $\widehat{q}$, whose cost bounds $\mathrm{EMD}(p,\widehat{q})$ from above. The second step gives a feasible plan from q_a to q_f, whose cost bounds $\mathrm{EMD}(q_a,q_f)$ from above.

Since

$$\mathrm{EMD}(\mu,q)\le\mathrm{EMD}(\mu,\widehat{q})+\mathrm{EMD}(\widehat{q},q)=\mathrm{EMD}(\mu,\widehat{q})+\mathrm{EMD}(q_a,q_f),$$

it suffices to show that

$$\mathrm{EMD}(\mu,\widehat{q})\le(1+O(\varepsilon))\,\mathrm{EMD}(p,\widehat{q})+O(\varepsilon r).$$

Construct an auxiliary graph G on $\mathrm{supp}(q)\cup A$ as follows. For every $i\in[t]$ and $y\in\mathrm{supp}(q)$ with $\rho(a_i,y)\le\varepsilon^{-1}r$, we add an edge between a_i and y, i.e. we connect every close pair. Let $\mathcal{S}$ denote the partition determined by the connected components of G. For each connected component $S\in\mathcal{S}$, we have that

$$p(\bigcup_{i:a_i\in S}B_i)=p_c(\bigcup_{i:a_i\in S}B_i)+p_f(\bigcup_{i:a_i\in S}B_i)=p_c(\bigcup_{i:a_i\in S}B_i)+\sum_{i:a_i\in S}q_a(a_i).$$

Since every close transportation corresponds to an edge of G and S is a connected component, all close transportations incident to $a_i\in S$ occur only within S. Thus, $p_c(\bigcup_{i:a_i\in S}B_i)=q_c(\mathrm{supp}(q)\cap S)$ and therefore

$$p(\bigcup_{i:a_i\in S}B_i)=\widehat{q}(S).$$

By the construction of $\mathcal{N}$, there exists a distribution $\tilde{q}\in\mathcal{N}$ such that (1) $\mathrm{supp}(\tilde{q})=\mathrm{supp}(\widehat{q})$; (2) the corresponding partition of $\tilde{q}$ is exactly $\mathcal{S}$; and (3) $\sum_{y\in\mathrm{supp}(q)}|\tilde{q}(y)-\widehat{q}(y)|\le 2\varepsilon^2/k$. We claim that $\mathrm{EMD}(\tilde{q},\widehat{q})\le O(\varepsilon r)$. Indeed, on each connected component of G, $\tilde{q}$ and $\widehat{q}$ have the same total mass. Thus,

$$\mathrm{EMD}(\tilde{q},\widehat{q})\le\frac{2kr}{\varepsilon}\cdot\frac{2\varepsilon^2}{k}=O(\varepsilon r),$$

since the distance between two points in the same connected component can be bounded by $2kr/\varepsilon$.

Next, by the coreset property of $\tilde{q}\in\mathcal{N}$, $\mathrm{EMD}(\mu,\tilde{q})\le(1+\varepsilon)\,\mathrm{EMD}(p,\tilde{q})+\varepsilon r$. Therefore,

$$\begin{aligned}\mathrm{EMD}(\mu,\widehat{q})&\le\mathrm{EMD}(\mu,\tilde{q})+\mathrm{EMD}(\tilde{q},\widehat{q})\\&\le(1+\varepsilon)\,\mathrm{EMD}(p,\tilde{q})+O(\varepsilon r)\\&\le(1+\varepsilon)(\mathrm{EMD}(p,\widehat{q})+\mathrm{EMD}(\widehat{q},\tilde{q}))+O(\varepsilon r)\\&\le(1+\varepsilon)\,\mathrm{EMD}(p,\widehat{q})+O(\varepsilon r).\end{aligned}$$

□

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract emphasizes the linearity in k and the dependence on additional parameters is clearly stated in Theorem 3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [No]

Justification: This is a theoretical paper, and all assumptions and results have been stated clearly and formally. Our main theorem is presented in full generality and can be applied to an arbitrary dataset. Thus, we decided not to include further discussions on limitations in the main text.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: This is a theoretical paper and all assumptions of the theorems and corollaries are stated clearly. Proofs omitted in the main body are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper is theoretical in nature, no experiments are involved and no societal impacts are expected.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper is theoretical in nature and no societal impacts are expected.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper is theoretical in nature and poses no misuse risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLM is used only for editing purposes.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.