k-variates++: more pluses in the k-means++ — Supplementary Information —

Richard Nock
Data61 & The Australian National University
richard.nock@data61.csiro.au

Raphaël Canyasse
Ecole Polytechnique & The Technion
raphael.canyasse@polytechnique.edu
Raphael.can@tx.technion.ac.il

Roksana Boreli
Data61 & The University of New South Wales
r.boreli@unsw.edu.au

Frank Nielsen
Ecole Polytechnique & Sony Computer Science Laboratories, Inc.
Frank.Nielsen@acm.org

May 25, 2016

Abstract

This is the Supplementary Information to Paper "k-variates++: more pluses in the k-means++", appearing in the proceedings of ICML 2016. Notation "main file" indicates reference to the paper.

1 Table of contents

2 Supplementary Material on Proofs

Several proofs rely on properties of the k-means++ algorithm that are not exploited in the proof of [1]. We assume here the basic knowledge of the proof technique of [1].

2.1 Proof of Theorem 2

Let A denote a subset of $\mathcal{A}$, and $\boldsymbol{c}(A) \doteq (1/|A|) \cdot \sum_{\boldsymbol{a} \in A} \boldsymbol{a}$ the barycenter of A. It is well known that $\boldsymbol{c}(A) = \arg\min_{\boldsymbol{a}' \in \mathbb{R}^d} \sum_{\boldsymbol{a} \in A} \|\boldsymbol{a} - \boldsymbol{a}'\|_2^2$, so the potential of A,

$$\phi(A) \doteq \sum_{\boldsymbol{a} \in A} \|\boldsymbol{a} - \boldsymbol{c}(A)\|_2^2 \qquad (1)$$

is just the optimal potential of A if A defines a cluster in the optimal clustering. We also define the noisy potential of A as:

$$\phi^N(A) \doteq \sum_{\boldsymbol{a} \in A} \int_{\Omega_{\boldsymbol{a}}} \|\boldsymbol{x} - \boldsymbol{c}(A)\|_2^2 \mathrm{d}p_{\boldsymbol{a}}(\boldsymbol{x}) \ . \qquad (2)$$

The proof of Theorem 2 follows the same path as the proof of Theorem 3.1 in [1]. Instead of reproducing the proof, we shall assume basic knowledge of the original proof and will just provide the side Lemmata that are sufficient for our more general result. The first Lemma is a generalization of Lemma 3.2 in [1].

Lemma 1 *Let C_{opt} denotes the optimal partition of $\mathcal{A}$ according to eq. (2). Let A be an arbitrary cluster in C_{opt}. Let C be a single-cluster clustering whose center is chosen at random by one step of Algorithm k-variates++ (i.e. for $t = 1$). Then*

$$\mathbb{E}[\phi(A)] = \phi_{\mathrm{opt}}(A) + \phi^N_{\mathrm{opt}}(A) \ . \qquad (3)$$

Proof The expected potential of cluster A is

$$\begin{aligned}
&\mathbb{E}[\phi(A; \mathcal{C} = \emptyset)] \\
&= \frac{1}{|A|} \cdot \sum_{\boldsymbol{a}_0 \in A} \int_{\Omega_{\boldsymbol{a}_0}} \sum_{\boldsymbol{a} \in A} \|\boldsymbol{a} - \boldsymbol{x}\|_2^2 \mathrm{d}p_{\boldsymbol{a}_0}(\boldsymbol{x}) \\
&= \frac{1}{|A|} \cdot \sum_{\boldsymbol{a}_0 \in A} \int_{\Omega_{\boldsymbol{a}_0}} \sum_{\boldsymbol{a} \in A} \|\boldsymbol{a} - \boldsymbol{c}(A) + \boldsymbol{c}(A) - \boldsymbol{x}\|_2^2 \mathrm{d}p_{\boldsymbol{a}_0}(\boldsymbol{x}) \\
&= \frac{1}{|A|} \cdot \sum_{\boldsymbol{a}_0 \in A} \left(\begin{array}{l} \sum_{\boldsymbol{a} \in A} \|\boldsymbol{a} - \boldsymbol{c}(A)\|_2^2 + |A| \cdot \int_{\Omega_{\boldsymbol{a}_0}} \|\boldsymbol{x} - \boldsymbol{c}(A)\|_2^2 \mathrm{d}p_{\boldsymbol{a}_0}(\boldsymbol{x}) \\ +2 \sum_{\boldsymbol{a} \in A} \langle \boldsymbol{a} - \boldsymbol{c}(A), \boldsymbol{c}(A) - \int_{\Omega_{\boldsymbol{a}_0}} \boldsymbol{x} \mathrm{d}p_{\boldsymbol{a}_0}(\boldsymbol{x}) \rangle \end{array} \right) \\
&= \frac{1}{|A|} \cdot \sum_{\boldsymbol{a}_0 \in A} \left(\begin{array}{l} \sum_{\boldsymbol{a} \in A} \|\boldsymbol{a} - \boldsymbol{c}(A)\|_2^2 + |A| \cdot \int_{\Omega_{\boldsymbol{a}_0}} \|\boldsymbol{x} - \boldsymbol{c}(A)\|_2^2 \mathrm{d}p_{\boldsymbol{a}_0}(\boldsymbol{x}) \\ +2 \langle \underbrace{\sum_{\boldsymbol{a} \in A} \boldsymbol{a} - |A| \boldsymbol{c}(A)}_{=0}, \boldsymbol{c}(A) - \boldsymbol{a}_0 \rangle \end{array} \right) \\
&= \sum_{\boldsymbol{a} \in A} \|\boldsymbol{a} - \boldsymbol{c}(A)\|_2^2 + \sum_{\boldsymbol{a} \in A} \int_{\Omega_{\boldsymbol{a}_0}} \|\boldsymbol{x} - \boldsymbol{c}(A)\|_2^2 \mathrm{d}p_{\boldsymbol{a}}(\boldsymbol{x}) \\
&= \phi_{\mathrm{opt}}(A) + \phi^N_{\mathrm{opt}}(A) \ ,
\end{aligned}$$

as claimed. ■

When $p_{\boldsymbol{a}}$ is a Dirac anchored at $\boldsymbol{a}$, we recover Lemma 3.2 in [1]. The following Lemma generalizes Lemma 3.3 in [1].

Lemma 2 *Suppose that the optimal clustering C_{opt} is η-probe approximable. Let A be an arbitrary cluster in C_{opt}, and let C be an arbitrary clustering with centers $\mathcal{C}$. Suppose that the reference point $\boldsymbol{a}$ chosen according to (1) in Step 2.1 is in A. Then the random point $\boldsymbol{x}$ picked in Step 2.2 brings an expected potential that satisfies*

$$\mathbb{E}[\phi(A)] \quad \leq \quad (6+4\eta)\cdot\phi_{\mathrm{opt}}(A) + 2\cdot\phi^{N}_{\mathrm{opt}}(A) \ . \tag{4}$$

Proof Let us denote $\boldsymbol{c}^{\star}(\boldsymbol{u}) \doteq \arg\min_{\boldsymbol{x}\in\mathcal{C}} \|\boldsymbol{u}-\boldsymbol{x}\|_2^2$ (since $C \neq C_{\mathrm{opt}}$ in general, $\boldsymbol{c}^{\star}(\boldsymbol{u}) \neq \boldsymbol{c}_{\mathrm{opt}}(\boldsymbol{u})$), and $D(\boldsymbol{a}) \doteq \|\boldsymbol{a}-\boldsymbol{c}^{\star}(\boldsymbol{a})\|_2^2$ the contribution of $\boldsymbol{a}\in A$ to the k-means potential defined by $\mathcal{C}$. We have, using Lemma 3.3 in [1] and Lemma 1,

$$\mathbb{E}_{\boldsymbol{x}}[\phi(A;\mathcal{C}\cup\{\boldsymbol{x}\})] \quad = \quad \sum_{\boldsymbol{a}_0\in A}\frac{D_t(\boldsymbol{a}_0)}{\sum_{\boldsymbol{a}\in A}D_t(\boldsymbol{a})}\cdot\sum_{\boldsymbol{a}\in A}\int_{\Omega_{\boldsymbol{a}_0}}\min\{D(\boldsymbol{a}),\|\boldsymbol{a}-\boldsymbol{x}\|_2^2\}\mathrm{d}p_{\boldsymbol{a}_0}(\boldsymbol{x}) \ . \tag{5}$$

The triangle inequality gives, for any $\boldsymbol{a}\in A$,

$$\begin{aligned}\sqrt{D_t(\boldsymbol{a}_0)} \quad &\doteq \quad \|\wp_t(\boldsymbol{a}_0)-\boldsymbol{c}^{\star}(\wp_t(\boldsymbol{a}_0))\|_2\\ &\leq \quad \|\wp_t(\boldsymbol{a}_0)-\boldsymbol{c}^{\star}(\wp_t(\boldsymbol{a}))\|_2\\ &\leq \quad \|\wp_t(\boldsymbol{a}_0)-\wp_t(\boldsymbol{a})\|_2+\|\wp_t(\boldsymbol{a})-\boldsymbol{c}^{\star}(\wp_t(\boldsymbol{a}))\|_2 \ ;\end{aligned} \tag{6}$$

since $(a+b)^2\leq 2a^2+2b^2$, then $D_t(\boldsymbol{a}_0)\leq 2\|\wp_t(\boldsymbol{a}_0)-\wp_t(\boldsymbol{a})\|_2^2+2D_t(\boldsymbol{a})$, and so, after averaging over A,

$$D_t(\boldsymbol{a}_0) \quad \leq \quad \frac{2}{|A|}\sum_{\boldsymbol{a}\in A}\|\wp_t(\boldsymbol{a}_0)-\wp_t(\boldsymbol{a})\|_2^2+\frac{2}{|A|}\sum_{\boldsymbol{a}\in A}D_t(\boldsymbol{a}) \ , \tag{7}$$

and eq. (5) can be upperbounded as:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[\phi(A;\mathcal{C}\cup\{\boldsymbol{x}\})] \quad &\leq \quad \frac{2}{|A|}\sum_{\boldsymbol{a}_0\in A}\frac{\sum_{\boldsymbol{a}\in A}\|\wp_t(\boldsymbol{a}_0)-\wp_t(\boldsymbol{a})\|_2^2}{\sum_{\boldsymbol{a}\in A}D_t(\boldsymbol{a})}\cdot\sum_{\boldsymbol{a}\in A}\int_{\Omega_{\boldsymbol{a}_0}}\min\{D(\boldsymbol{a}),\|\boldsymbol{a}-\boldsymbol{x}\|_2^2\}\mathrm{d}p_{\boldsymbol{a}_0}(\boldsymbol{x})\\ &\quad +\frac{2}{|A|}\sum_{\boldsymbol{a}_0\in A}\frac{\sum_{\boldsymbol{a}\in A}D_t(\boldsymbol{a})}{\sum_{\boldsymbol{a}\in A}D_t(\boldsymbol{a})}\cdot\sum_{\boldsymbol{a}\in A}\int_{\Omega_{\boldsymbol{a}_0}}\min\{D(\boldsymbol{a}),\|\boldsymbol{a}-\boldsymbol{x}\|_2^2\}\mathrm{d}p_{\boldsymbol{a}_0}(\boldsymbol{x})\\ &\leq \quad \underbrace{\frac{2}{|A|}\sum_{\boldsymbol{a}_0\in A}\frac{\sum_{\boldsymbol{a}\in A}D(\boldsymbol{a})}{\sum_{\boldsymbol{a}\in A}D_t(\boldsymbol{a})}\cdot\sum_{\boldsymbol{a}\in A}\|\wp_t(\boldsymbol{a}_0)-\wp_t(\boldsymbol{a})\|_2^2}_{\doteq P_1}\\ &\quad \underbrace{+\frac{2}{|A|}\sum_{\boldsymbol{a}_0\in A}\sum_{\boldsymbol{a}\in A}\int_{\Omega_{\boldsymbol{a}_0}}\|\boldsymbol{a}-\boldsymbol{x}\|_2^2\mathrm{d}p_{\boldsymbol{a}_0}(\boldsymbol{x})}_{\doteq P_2} \ .\end{aligned} \tag{8}$$

We bound the two potentials P_1 and P_2 separately, starting with P_1. Fix any $\boldsymbol{a}_0 \in A$. If $\sum_{\boldsymbol{a}\in A} \|\wp_t(\boldsymbol{a}) - \wp_t(\boldsymbol{a}_0)\|_2^2 = 0$, then trivially

$$\left(\sum_{\boldsymbol{a}\in A} D(\boldsymbol{a})\right) \cdot \left(\sum_{\boldsymbol{a}\in A} \|\wp_t(\boldsymbol{a}_0) - \wp_t(\boldsymbol{a})\|_2^2\right) \leq (1+\eta) \cdot \left(\sum_{\boldsymbol{a}\in A} D_t(\boldsymbol{a})\right) \cdot \left(\sum_{\boldsymbol{a}\in A} \|\boldsymbol{a}_0 - \boldsymbol{a}\|_2^2\right) \quad (9)$$

since the right-hand side cannot be negative. If $\sum_{\boldsymbol{a}\in A} \|\wp_t(\boldsymbol{a}) - \wp_t(\boldsymbol{a}_0)\|_2^2 \neq 0$, then since $\wp_t$ is η-stretching, we have:

$$\frac{\sum_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \boldsymbol{c}^\star(\boldsymbol{a})\|_2^2}{\sum_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \boldsymbol{a}_0\|_2^2} \leq (1+\eta) \cdot \frac{\sum_{\boldsymbol{a}\in A} \|\wp_t(\boldsymbol{a}) - \boldsymbol{c}^\star(\wp_t(\boldsymbol{a}))\|_2^2}{\sum_{\boldsymbol{a}\in A} \|\wp_t(\boldsymbol{a}) - \wp_t(\boldsymbol{a}_0)\|_2^2} , \quad (10)$$

which is exactly ineq. (9) after rearranging the terms. Ineq (9) implies

$$\begin{aligned} P_1 &\leq 2(1+\eta) \cdot \frac{1}{|A|} \sum_{\boldsymbol{a}_0\in A} \sum_{\boldsymbol{a}\in A} \|\boldsymbol{a}_0 - \boldsymbol{a}\|_2^2 \\ &= 4(1+\eta) \cdot \phi_{\mathrm{opt}}(A) , \end{aligned} \quad (11)$$

where the equality follows from [1], Lemma 3.2. Also, Lemma 1 brings

$$\begin{aligned} P_2 &= 2 \cdot \frac{1}{|A|} \sum_{\boldsymbol{a}_0\in A} \int_{\Omega_{\boldsymbol{a}_0}} \sum_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \boldsymbol{x}\|_2^2 \mathrm{d}p_{\boldsymbol{a}_0}(\boldsymbol{x}) \\ &= 2\phi_{\mathrm{opt}}(A) + 2\phi^N_{\mathrm{opt}}(A) . \end{aligned} \quad (12)$$

We therefore get

$$\mathbb{E}_{\boldsymbol{x}}[\phi(A; \mathcal{C} \cup \{\boldsymbol{x}\})] \leq (6+4\eta) \cdot \phi_{\mathrm{opt}}(A) + 2 \cdot \phi^N_{\mathrm{opt}}(A) , \quad (13)$$

as claimed. ■

Again, we recover Lemma 3.3 in [1] when $p_{\boldsymbol{a}}$ is a Dirac and the probe function $\wp = \mathrm{Id}$. The rest of the proof of Theorem 2 consists of the same steps as Theorem 3.1 in [1], after having remarked that $\phi^N_{\mathrm{opt}}(A)$ can be simplified:

$$\begin{aligned} \phi^N_{\mathrm{opt}}(A) &= \sum_{\boldsymbol{a}\in A} \int_{\Omega_{\boldsymbol{a}_0}} \|\boldsymbol{x} - \boldsymbol{c}(A)\|_2^2 \mathrm{d}p_{\boldsymbol{a}}(\boldsymbol{x}) \\ &= \sum_{\boldsymbol{a}\in A} \int_{\Omega_{\boldsymbol{a}_0}} \|\boldsymbol{x}\|_2^2 \mathrm{d}p_{\boldsymbol{a}}(\boldsymbol{x}) - 2\langle \boldsymbol{c}(A), \boldsymbol{\mu}_{\boldsymbol{a}}\rangle + \|\boldsymbol{c}(A)\|_2^2 \\ &= \sum_{\boldsymbol{a}\in A} \int_{\Omega_{\boldsymbol{a}_0}} \|\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{a}}\|_2^2 \mathrm{d}p_{\boldsymbol{a}}(\boldsymbol{x}) + \|\boldsymbol{\mu}_{\boldsymbol{a}}\|_2^2 - 2\langle \boldsymbol{c}(A), \boldsymbol{a}\rangle + \|\boldsymbol{c}(A)\|_2^2 \\ &= \sum_{\boldsymbol{a}\in A} \left\{ \mathrm{tr}\left(\Sigma_{\boldsymbol{a}}\right) + \|\boldsymbol{\mu}_{\boldsymbol{a}} - \boldsymbol{c}(A)\|_2^2 \right\} \\ &= \phi_{\mathrm{bias}}(A) + \phi_{\mathrm{var}}(A) . \end{aligned} \quad (14)$$

2.2 Proof of Lemma 3

The proof is a simple application of the Fréchet-Cramér-Rao-Darmois bound. Consider the simple case $k = 1$ and a spherical Gaussian noise for p with a single point in $\mathcal{A}$. Renormalize both sides of (7) by $m \doteq |\mathcal{A}|$ so that $(1/m)\sum_{a\in\mathcal{A}} \mathrm{tr}\left(\Sigma_{\boldsymbol{a}}\right) = \mathrm{tr}\left(\Sigma\right)$. One sees that the left hand side of ineq. (7) is just an estimator of the variance of $p_{\boldsymbol{a}}$, which, by Fréchet-Darmois-Cramér-Rao bound, has to be at least the inverse of the Fisher information, that is in this case, the trace of the covariance matrix, *i.e.* $\mathrm{tr}\left(\Sigma\right)$.

2.3 Extension and comments on Table 1

Before embarking on the proofs, Table 1 below provides a more extensive comparison to the state of the art in distributed, streamed and on-line clustering than Table 1 in the main file. Notation O^* removes all dependencies in their model parameters (assumptions, model parameters, and δ for the (ϵ,δ)-DP in [2]), and λ is the separability assumption parameter [3][1]. The approximation bounds in [3] consider Wasserstein distance between (estimated / optimal) centers, and not the potential involving data points like us. To obtain bounds that can be compared, we have used the simple trick that the observed potential is, up to a constant, no more than the optimal potential plus a fonction of the distance between (estimated / optimal) centers. This somewhat degrades the bound, but not enough for the observed discrepancies with our bound to reverse or even vanish. It is clear from the bounds that the noise dependence is significantly in our favor, and our bound is also significantly better at least when k is not too large. To be a bit more specific, [2] are concerned with approximating subspace clustering, and so they are using a very different potential function, which is, between two subspaces $\mathcal{S}$ and $\mathcal{S}'$, $d(\mathcal{S},\mathcal{S}') = \|\mathrm{UU}^\top - \mathrm{U}'\mathrm{U}'^\top\|_F$, where U (resp. U') is an *orthonormal* basis for $\mathcal{S}$ (resp. $\mathcal{S}'$). To obtain an idea of the approximation on the k-means clustering problem that their technique yields, we compute ϕ in the projected space, using the fact that, because of the triangle inequality and the fact that projections are linear and do not increase norms,

$$\begin{aligned}
\|\mathrm{proj}_{\mathrm{U}}(\boldsymbol{a}) - \mathrm{proj}_{\mathrm{U}'}(\boldsymbol{a}')\|_2 &= \|(\mathrm{proj}_{\mathrm{U}}(\boldsymbol{a}) - \mathrm{proj}_{\mathrm{U}}(\boldsymbol{a}')) + (\mathrm{proj}_{\mathrm{U}}(\boldsymbol{a}') - \mathrm{proj}_{\mathrm{U}'}(\boldsymbol{a}'))\|_2 \quad (15)\\
&\leq \|\mathrm{proj}_{\mathrm{U}}(\boldsymbol{a}) - \mathrm{proj}_{\mathrm{U}}(\boldsymbol{a}')\|_2 + \|\mathrm{proj}_{\mathrm{U}}(\boldsymbol{a}') - \mathrm{proj}_{\mathrm{U}'}(\boldsymbol{a}'))\|_2 \quad (16)\\
&\leq \|\mathrm{proj}_{\mathrm{U}}(\boldsymbol{a}) - \mathrm{proj}_{\mathrm{U}}(\boldsymbol{a}')\|_2 + 2\|\boldsymbol{a}'\|_2 \ . \quad (17)
\end{aligned}$$

To account for the approximation in the inequalities, we then discard the rightmost term, replacing therefore $\|\mathrm{proj}_{\mathrm{U}}(\boldsymbol{a}) - \mathrm{proj}_{\mathrm{U}'}(\boldsymbol{a}')\|_2$ by $\|\mathrm{proj}_{\mathrm{U}}(\boldsymbol{a}) - \mathrm{proj}_{\mathrm{U}}(\boldsymbol{a}')\|_2$, which amounts, in the approximation bounds, to remove the dependence in the dimension. At this price, and using the trick to transfer the wasserstein distance between centers to L_2^2 potential between points to cluster centers, we obtain the approximation bound in (β) of Table 1. While it has to be used with care, its main interest is in showing that the price to pay because of the noise component is in fact not decreasing in m.

[1] λ is named ϕ in [3]. We use λ to avoid confusion with clustering potentials.

	Ref.	Property	Them	Us
(1)	[4]	Communication complexity	$O(n^2\ell \cdot \log \phi_1)$ (expected)	$O(n^2k)$
(2)	[4]	# data to compute one center	m	$\leq \max_{i\in[n]}(m/m_i)$
(3)	[4]	Data points shared	$O(\ell \cdot \log \phi_1)$ (expected)	k
(4)	[4]	Approximation bound	$O((\log k) \cdot \phi_{\mathrm{opt}})$	$(2+\log k) \cdot \left(10\phi_{\mathrm{opt}} + 6\phi_s^F\right)$
(I)	[5]	Communication complexity	$\Omega((nkd/\varepsilon^4) + n^2k\ln(nk))$	$O(n^2k)$
(II)	[5]	Data points shared	$\Omega((kd/\varepsilon^4) + nk\ln(nk))$	k
(III)	[5]	Approximation bound	$(2+\log k)(1+\varepsilon) \cdot 8\phi_{\mathrm{opt}}$	$(2+\log k) \cdot \left(10\phi_{\mathrm{opt}} + 6\phi_s^F\right)$
(i)	[6]	Time complexity (outer loop)	— identical —	
(ii)	[6]	Approximation bound	$(2+\log k)(1+\eta) \cdot 32\phi_{\mathrm{opt}}$	$(2+\log k) \cdot ((8+4\eta)\phi_{\mathrm{opt}} + 2\phi_s^{\wp})$
(a)	[7]	Knowledge required	Lowerbound $\phi^* \leq \phi_{\mathrm{opt}}$	None
(b)	[7]	Approximation bound	$O(\log m \cdot \phi_{\mathrm{opt}})$	$(2+\log k) \cdot \left(4 + (32/\varsigma^2)\right) \phi_{\mathrm{opt}}$
(A)	[3]	Knowledge required	$\lambda(\phi_{\mathrm{opt}})$	None
(B)	[3]	Noise variance (σ)	$O(\lambda kR/\epsilon)$	$O(R/(\epsilon + \log m))$
(C)	[3]	Approximation bound	$O^*(\phi_{\mathrm{opt}} + m\lambda^2kR^2/\epsilon^2)$	$O(\log k(\phi_{\mathrm{opt}} + mR^2/(\epsilon + \log m)^2))$
(α)	[2]	Assumptions on ϕ_{opt}	Several (separability, size of clusters, etc.)	None
(β)	[2]	Approximation bound	$O^*(\phi_{\mathrm{opt}} + km\log(m)R^2/\epsilon^2)$	$O(\log k(\phi_{\mathrm{opt}} + mR^2/(\epsilon + \log m)^2))$

Table 1: Comparison with state of the art approaches for distributed clustering (1-4, I-III), streamed clustering (i, ii), on-line clustering (a, b) and differential privacy (A-C, α, β). Notations used for the "Them" column are as follows. ϕ_1 is the expected potential of a clustering with a single cluster over the *whole* data and ℓ is in general $\Omega(k)$ [4]. ε is the coreset approximation factor in [5]. η is the approximation factor of the optimum in [6]. λ is the separability factor in Definition 5.1 in [3].

2.4 Proofs of Theorems 4, 5 and 6

The proof of these Theorems uses a *reduction* from k-variates++ to the corresponding algorithms, meaning that there exists particular probe functions and densities for which the set of centers delivered by k-variates++ is the same as the one delivered by the corresponding algorithms.

Definition 3 *Let $\mathcal{H}$ (parameters omitted) be any hard membership k-clustering algorithm. We way that k-variates++ **reduces** to $\mathcal{H}$ iff there exists data, densities and probe functions depending on the instance of $\mathcal{H}$ such that, in expectation over the internal randomisation of $\mathcal{H}$, the set of centers delivered by $\mathcal{H}$ are the same as the ones delivered by k-variates++. We note it*

$$k\text{-variates++} \;\; \succeq \;\; \mathcal{H} \;. \tag{18}$$

Hence, whenever k-variates++ $\succeq \mathcal{H}$, Theorem 2 in the main file immediately gives a guarantee for the approximation of the global optimum in expectation for $\mathcal{H}$, but this requires the translation of the parameters involved in Φ in ineq. (7) in the main file to involve only parameters from $\mathcal{H}$. In all our examples, this translation poses no problem at all.

Proof of Theorem 4

To have a concrete idea of one setting in which we can run and analyze Dk-means++/PDk-means++, we consider a privacy setting for the analysis. As already sketched in the main paper, Figure 1 presents the architecture of message passing in the Dk-means++/PDk-means++ framework. We first focus on the protected scheme, Dk-means++. We reduce k-variates++ to Algorithm 1 using identity probe functions: $\wp_t = \mathrm{Id}, \forall t$. The trick in reduction relies on the densities. We let $p_{\boldsymbol{\mu_a},\boldsymbol{\theta_a}}$ be uniform over the subset $\mathcal{A}_i$ to which $\boldsymbol{a}$ belongs. Thus, the support of densities is discrete,

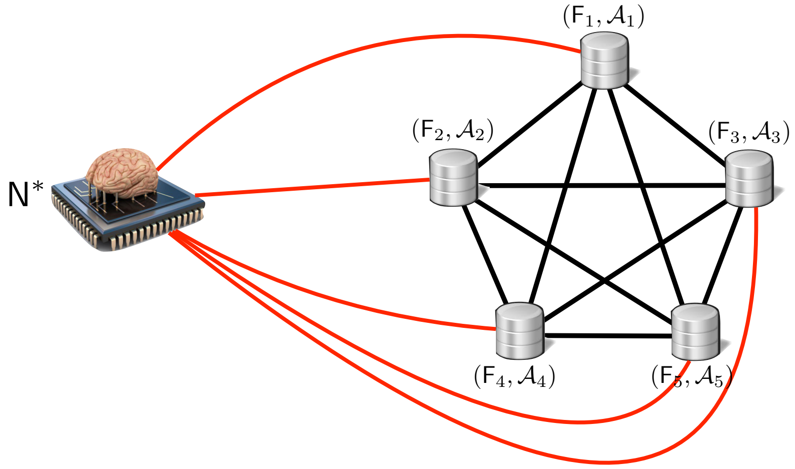

$(\mathsf{F}_1, \mathcal{A}_1)$
$(\mathsf{F}_2, \mathcal{A}_2)$
$(\mathsf{F}_3, \mathcal{A}_3)$
N*
$(\mathsf{F}_4, \mathcal{A}_4)$
$(\mathsf{F}_5, \mathcal{A}_5)$

Figure 1: Message passing between peers / nodes in the Dk-means++/PDk-means++ framework. Black edges and red arcs denote message passing between peers / nodes. On each black edge circulates at most k data points; on each red arcs circulates k total potentials.

and $\mathcal{C}$ is a subset of $\mathcal{A}$; furthermore, the probability $q_t(\boldsymbol{a})$ that $\boldsymbol{a} \in \mathcal{A}_i$ is chosen at iteration t in k-variates++ actually simplifies to a convenient expression:

$$q_t(\boldsymbol{a}) \quad = \quad q^D_{ti} \cdot u_i \quad , \tag{19}$$

where we recall that

$$q^D_{ti} \quad \doteq \quad \left\{ \begin{array}{ccl} D_t(\mathcal{A}_i) \cdot (\sum_j D_t(\mathcal{A}_j))^{-1} & \text{if} & t > 1 \\ (1/n) & \text{otherwise} & \end{array} \right. . \tag{20}$$

Hence, picking $\boldsymbol{a}$ can be equivalently done by first picking $\mathcal{A}_i$ using q^D_t, and then, given the i chosen, sampling uniformly at random $\boldsymbol{a}$ in $\mathcal{A}_i$, which is what Forgy nodes do. We therefore get the equivalence between Algorithm 1 and k-variates++ as instantiated.

Lemma 4 *With data, densities and probes defined as before, k-variates++ $\succeq$ Dk-means++.*

To get the approximability ratio of Dk-means++, we translate the parameters of Φ in ineq. (7) in the main file. First, since $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\begin{aligned}
\phi_{\mathrm{bias}} &\doteq \sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{\mu}_{\boldsymbol{a}} - \boldsymbol{c}_{\mathrm{opt}}(\boldsymbol{a})\|_2^2 \\
&= \sum_{i\in[n]} \sum_{\boldsymbol{a}\in\mathcal{A}_i} \|\boldsymbol{c}(\mathcal{A}_i) - \boldsymbol{c}_{\mathrm{opt}}(\boldsymbol{a})\|_2^2 \qquad (21)\\
&= \sum_{i\in[n]} \sum_{\boldsymbol{a}\in\mathcal{A}_i} \|\boldsymbol{c}(\mathcal{A}_i) - \boldsymbol{a} + \boldsymbol{a} - \boldsymbol{c}_{\mathrm{opt}}(\boldsymbol{a})\|_2^2 \\
&\leq 2\sum_{i\in[n]} \sum_{\boldsymbol{a}\in\mathcal{A}_i} \|\boldsymbol{c}(\mathcal{A}_i) - \boldsymbol{a}\|_2^2 + 2\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \boldsymbol{c}_{\mathrm{opt}}(\boldsymbol{a})\|_2^2 \\
&= 2\phi^F_s + 2\phi_{\mathrm{opt}} \ . \qquad (22)
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\phi_{\mathrm{var}} &\doteq \sum_{\boldsymbol{a}\in\mathcal{A}} \mathrm{tr}\left(\Sigma_{\boldsymbol{a}}\right) \\
&= \sum_{\boldsymbol{a}\in\mathcal{A}} \int_{\Omega_{\boldsymbol{a}}} \|\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{a}}\|_2^2 \mathrm{d}p_{\boldsymbol{a}}(\boldsymbol{x}) \\
&= \sum_{i\in[n]} \sum_{\boldsymbol{a}\in\mathcal{A}_i} \sum_{\boldsymbol{a}'\in\mathcal{A}_i} \frac{1}{m_i} \cdot \|\boldsymbol{a}' - \boldsymbol{c}(\mathcal{A}_i)\|_2^2 \\
&= \sum_{i\in[n]} \sum_{\boldsymbol{a}\in\mathcal{A}_i} \|\boldsymbol{a} - \boldsymbol{c}(\mathcal{A}_i)\|_2^2 = \phi^F_s \ . \qquad (23)
\end{aligned}$$

There remains to plug ineq. (22) and eq. (23) in Theorem 2, along with $\eta = 0$ (since $\wp = \mathrm{Id}$), to get $\mathbb{E}[\phi(\mathcal{A}; \mathcal{C})] \leq (2 + \log k) \cdot (10\phi_{\mathrm{opt}} + 6\phi_s)$, as in Theorem 4.

The private version, PDk-means++, follows immediately by leaving ϕ_{var} in Φ instead of carrying eq. (23). This ends the proof of Theorem 4.

Proof of Theorem 5

The proof proceeds in the same way as for Theorem 4. The probe function (the same for every iteration, $\wp_t = \wp, \forall t$) is already defined in the statement of Theorem 5, from the definition of synopses. The distributions $p_{\boldsymbol{\mu_a},\boldsymbol{\theta_a}}$ are Diracs anchored at the *probe* (synopses) locations. The centers chosen in k-variates++ are thus synopses, and it is not hard to check that the probability to pick a synopsis $\boldsymbol{s}_j$ at iteration t factors in the same way as in the definition of q^S_t in eq. (9) (main file). We therefore get the equivalence between Algorithm 2 and k-variates++ as instantiated.

Lemma 5 *With data, densities and probes defined as before, k-variates++ $\succeq$ Sk-means++.*

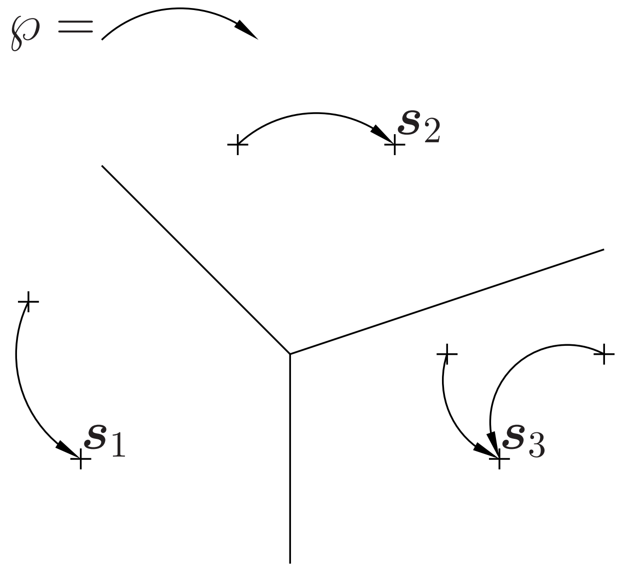


Figure 2: Computation of the probe function $\wp$ for the reduction from k-variates++ to $\mathrm{S}k$-means++. Segments display parts of the Voronoi diagram of $\mathcal{S}$.

The proof of the approximation property of $\mathrm{S}k$-means++ then follows from the fact that $\phi_{\mathrm{var}} = 0$ (Diracs) and

$$
\begin{aligned}
\phi_{\mathrm{bias}} &\doteq \sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{\mu_a} - \boldsymbol{c}_{\mathrm{opt}}(\boldsymbol{a})\|_2^2 \\
&= \sum_{\boldsymbol{a}\in\mathcal{A}} \|\wp(\boldsymbol{a}) - \boldsymbol{c}_{\mathrm{opt}}(\boldsymbol{a})\|_2^2 \\
&= \sum_{\boldsymbol{a}\in\mathcal{A}} \|\wp(\boldsymbol{a}) - \boldsymbol{a} + \boldsymbol{a} - \boldsymbol{c}_{\mathrm{opt}}(\boldsymbol{a})\|_2^2 \\
&\leq 2\sum_{\boldsymbol{a}\in\mathcal{A}} \|\wp(\boldsymbol{a}) - \boldsymbol{a}\|_2^2 + 2\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \boldsymbol{c}_{\mathrm{opt}}(\boldsymbol{a})\|_2^2 \\
&= 2\sum_{\boldsymbol{a}\in\mathsf{S}} \|\wp(\boldsymbol{a}) - \boldsymbol{a}\|_2^2 + 2\sum_{\boldsymbol{a}\in\mathsf{S}} \|\boldsymbol{a} - \boldsymbol{c}_{\mathrm{opt}}(\boldsymbol{a})\|_2^2 = 2\phi_s^{\wp} + 2\phi_{\mathrm{opt}} \qquad (24)
\end{aligned}
$$

(using again $(a+b)^2 \leq 2a^2 + 2b^2$). Using Theorem 2, this brings the statement of the Theorem.

Figure 2 shows that the "quality" of the probe function (spread $\phi_s^{\wp}$, stretching factor η) stem from the quality of the Voronoi diagram induced by the synopses in $\mathcal{S}$.

Proof of Theorem 6

The proof proceeds in the same way as for Theorem 4. The the reduction from k-variates++ to $\mathrm{OL}k$-means++ relies on two things: first, the uniform choice of the first center in k-means++ can be replaced by picking the center uniformly in *any* subset of the data: it does not change the expected approximation properties of the algorithm (this comes from Lemma 3.4 in [1]); therefore, the choice $q_1 \doteq u_m$ in k-variates++ can be replaced with $q_1 \doteq u_1$ (uniform with support $\mathcal{A}_1$). Second,

Setting	Algorithm	Probe functions $\wp_t$	Densities $p_{(\boldsymbol{\mu}_.,\boldsymbol{\theta}_.)}$
Batch	k-means++ [1]	Identity	Diracs
Distributed	Dk-means++	Identity	Uniform on data subsets
Distributed	PDk-means++	Identity	Non uniform, compact support
Streaming	Sk-means++	synopses	Diracs
On-line	OLk-means++	point (batch not hit) / closest center (batch hit)	Diracs

Table 2: Synthesis of the parameters for the reductions from k-variates++. We indicate k-means++ as the batch clustering solution [1].

a particular probe function needs to be devised, sketched in Figure 3. Basically, all probe functions of a minibatch are the same: each point in the minibatch is probed to itself, while points occurring outside the minibatch are probed to their closest center. The reduction proceeds in the following steps: we first let $\mathcal{A}$ be the complete set of points in the stream S. Then, we let $\mathcal{A}_j$ denote the set of points of minibatch S_j. Remark that minibatch $\mathcal{A}_j$ occurs in the stream before $\mathcal{A}_{j'}$ for $j < j'$, and minibatches induce a partition of $\mathcal{A}$. Let $j(t)$ denote the batch related to iteration t in k-variates++. We define the following probe function $\wp_t(\boldsymbol{a})$ in k-variates++, letting $\mathcal{A}_j$ the minibatch to which $\boldsymbol{a}$ belongs (we do not necessarily have $j = j(t)$):

- if $j = j(t)$, then $\wp_t(\boldsymbol{a}) \doteq \boldsymbol{a}$;
- else $\wp_t(\boldsymbol{a}) \doteq \arg\min_{\boldsymbol{c}\in\mathcal{C}} \|\boldsymbol{a} - \boldsymbol{c}\|_2^2$ (remark that $|\mathcal{C}| \geq 1$ in this case).

Finally, densities $p_{(\boldsymbol{\mu}_.,\boldsymbol{\theta}_.)}$ are Diracs anchored at selected points, like in k-means++. We get the equivalence between Algorithm 3 and k-variates++ as instantiated.

Lemma 6 *With data, densities and probes defined as before, k-variates++ $\succeq$ OLk-means++.*

The proof is immediate, since each minibatch is hit by a center exactly once in OLk-means++, and when one subset $\mathcal{A}_j$ is hit by a center, then the probe function makes that *no* other center can be sampled again from $\mathcal{A}_j$ (all contributions to the density q_t are then zero in $\mathcal{A}_j$). We now finish the proof of Theorem 6 by showing the same approximability ratio for k-variates++ as reduced. Because *optimal* clusters are ς-wide with respect to stream S, we have

$$\frac{1}{|A|} \cdot \sum_{\boldsymbol{a},\boldsymbol{a}' \in A} \|\boldsymbol{a} - \boldsymbol{a}'\|_2^2 \ \geq \ \varsigma \cdot R \ .$$

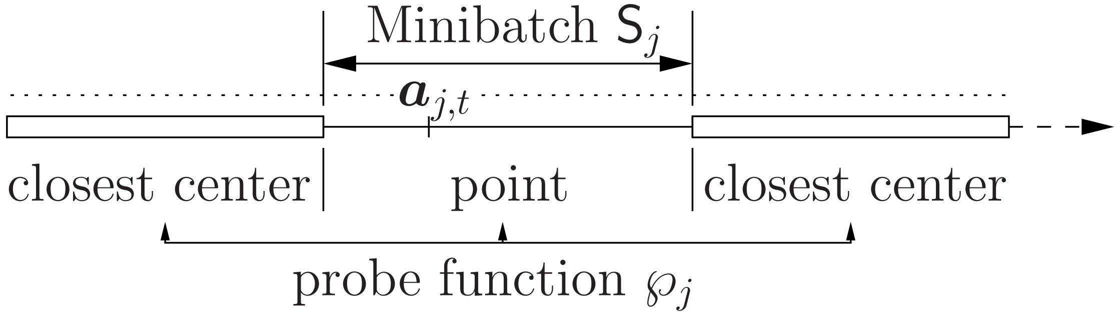


Figure 3: Computation of the probe function $\wp_t$ for the reduction from OLk-means++ to k-variates++, depending on each minibatch stream S_j.

Recall that $\boldsymbol{c}(A) \doteq (1/|A|) \cdot \sum_{\boldsymbol{a}\in A} \boldsymbol{a}$. For any $\boldsymbol{a}_0 \in A$, it holds that:

$$
\begin{aligned}
\frac{1}{|A|-1} \cdot \sum_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \boldsymbol{a}_0\|_2^2 &\geq \frac{1}{|A|-1} \cdot \sum_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \boldsymbol{c}(A)\|_2^2 && (25)\\
&= \frac{1}{|A|-1} \cdot \left(\frac{1}{2|A|} \cdot \sum_{\boldsymbol{a},\boldsymbol{a}'\in A} \|\boldsymbol{a} - \boldsymbol{a}'\|_2^2\right) && (26)\\
&= \frac{1}{4} \cdot \frac{2}{|A|(|A|-1)} \cdot \sum_{\boldsymbol{a},\boldsymbol{a}'\in A} \|\boldsymbol{a} - \boldsymbol{a}'\|_2^2 \\
&\geq \frac{\varsigma}{4} \cdot R\ . && (27)
\end{aligned}
$$

Ineq. (25) holds because $\boldsymbol{c}(A)$ is the population minimizer for optimal cluster A (see *e.g.*, [1], Lemma 2.1). Since probes are points of $\mathcal{A}$,

$$
\begin{aligned}
\phi(\wp_j(A); \{\wp_j(\boldsymbol{a}_0)\}) &\leq |A| \cdot R \\
&\leq \frac{4|A|}{\varsigma(|A|-1)} \cdot \sum_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \boldsymbol{a}_0\|_2^2\ . && (28)
\end{aligned}
$$

On the other hand, we have:

$$
\phi(\wp_t(A); \mathcal{C}) = \sum_{\boldsymbol{a}\in A\cap \mathsf{S}_j} \|\boldsymbol{a} - \boldsymbol{c}(\boldsymbol{a})\|_2^2\ , \qquad (29)
$$

but since minibatches are ς accurate, $\sum_{\boldsymbol{a}\in A\cap\mathsf{S}_j} \|\boldsymbol{a} - \boldsymbol{c}(\boldsymbol{a})\|_2^2 \geq \varsigma \cdot \sum_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \boldsymbol{c}(\boldsymbol{a})\|_2^2$. Therefore, for any $\boldsymbol{a}_0 \in A$,

$$
\begin{aligned}
\frac{\phi(\wp_t(A); \mathcal{C})}{\phi(\wp_t(A); \{\wp_t(\boldsymbol{a}_0)\})} &\geq \left(\frac{\varsigma^2(|A|-1)}{4|A|}\right) \cdot \frac{\sum_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \boldsymbol{c}(\boldsymbol{a})\|_2^2}{\sum_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \boldsymbol{a}_0\|_2^2} \\
&= \left(\frac{\varsigma^2(|A|-1)}{4|A|}\right) \cdot \frac{\phi(A; \mathcal{C})}{\phi(A; \{\boldsymbol{a}_0\})}\ . && (30)
\end{aligned}
$$

In other words, probe functions are η-stretching, for any η satisfying:

$$
\eta \geq \frac{4|A|}{\varsigma^2(|A|-1)} - 1\ , \qquad (31)
$$

and they are therefore η-stretching for $\eta = 8/\varsigma^2 - 1$. There remains to check that, because of the densities chosen,

$$\begin{aligned}\phi_{\mathrm{bias}} &= \phi_{\mathrm{opt}} \ , && (32)\\ \phi_{\mathrm{var}} &= 0 \ . && (33)\end{aligned}$$

This ends the proof of Theorem 6.

2.5 Proof of Theorem 9

To simplify notations in the proof, we let $p_{\boldsymbol{a}}(\boldsymbol{x})$ denote the value of density $p_{(\boldsymbol{\mu_a},\boldsymbol{\theta_a})}$ on some $\boldsymbol{x} \in \Omega$. Let us denote $Seq(n:k)$ the number of sequences of integers in set $\{1, 2, ..., n\}$ having exactly k elements, whose cardinal is $|Seq(n:k)| = n!/(n-k)!$. For any sequence $I \in Seq(n:k)$, we let I_i denote its i^{th} element. For any set $\mathcal{C} \doteq \{\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_k\}$ returned by Algorithm k-variates++with input instance set $\mathcal{A} \doteq \{\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n\} \subset \Omega$, the density of $\mathcal{C}$ given $\mathcal{A}$ is:

$$\mathbb{P}[\mathcal{C}|\mathcal{A}] = \sum_{\boldsymbol{\sigma}\in S_k}\sum_{I\in Seq(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A}) \ , \qquad (34)$$

where S_k denotes the symmetric group on k elements, and the following shorthand is used:

$$p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A}) \doteq \prod_{i=1}^{k} q_i(\boldsymbol{a}_{I_i}) p_{\boldsymbol{a}_{I_i}}(\boldsymbol{c}_{\sigma(i)}) \ , \qquad (35)$$

where q_i is computed using eq. (1) (main file) and taking into account the modification due to the choice of each I_j for $j < i$ in the sequence I.

In the following, we let $\mathcal{A}$ and $\mathcal{A}'$ denote two sets of points that differ from one a (they have the same size), say $\boldsymbol{a}_n \in \mathcal{A}$ and $\boldsymbol{a}'_n \in \mathcal{A}'$, $\boldsymbol{a}_n \neq \boldsymbol{a}'_n$. We analyze:

$$\frac{\mathbb{P}[\mathcal{C}|\mathcal{A}']}{\mathbb{P}[\mathcal{C}|\mathcal{A}]} = \frac{\sum_{\boldsymbol{\sigma}\in S_k}\sum_{I\in Seq(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A}')}{\sum_{\boldsymbol{\sigma}\in S_k}\sum_{I\in Seq(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A})} \ . \qquad (36)$$

Using the definition of $q(.)$, we refine $p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A})$ as

$$p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A}) = \frac{N(I)}{\prod_{i=1}^{k} M(I^i|\mathcal{A})} \cdot \prod_{i=1}^{k} p_{\boldsymbol{a}_{I_i}}(\boldsymbol{c}_{\sigma(i)}) \ , \qquad (37)$$

where

$$\begin{aligned} N(I) &\doteq \prod_{i=2}^{j} \|\boldsymbol{a}_{I_i} - \mathrm{NN}_{I^i}(\boldsymbol{a}_{I_i})\|_2^2 \ , && (38)\\ M(I^i|\mathcal{A}) &\doteq \left\{\begin{array}{cl} n & \text{if} \quad i=1 \\ \sum_{j=1}^{n} \|\boldsymbol{a}_j - \mathrm{NN}_{I^i}(\boldsymbol{a}_j)\|_2^2 & \text{otherwise} \end{array}\right. \ , && (39)\end{aligned}$$

and I^i is the prefix sequence $I_1, I_2, ..., I_{i-1}$, and $\mathrm{NN}_{I^i}(a) \doteq \arg\min_{j\leq i-1} \|a - \boldsymbol{a}_{I_j}\|_2$ is the nearest neighbor of a in the prefix sequence. Notice that there is a factor $1/m$ for $q(.)$ at the first iteration that we omit in $N(I)$ since it disappears in the ratio in eq. (36).

We analyze separately each element in (37), starting with $N(I)$. We define the *swapping* operation $s_\ell(I)$ that returns the sequence in which $\boldsymbol{a}_{I_\ell}$ and $\boldsymbol{a}_{I_{\ell+1}}$ are permuted, for $1 \leq \ell \leq k-1$. This incurs non-trivial modifications in $N(s_\ell(I))$ compared to $N(I)$, since the nearest neighbors of $\boldsymbol{a}_{I_\ell}$ and $\boldsymbol{a}_{I_{\ell+1}}$ may change in the permutation:

$$\begin{aligned} N(s_\ell(I)) &= \prod_{i=2}^{\ell-1} \|\boldsymbol{a}_{I_i} - \mathrm{NN}_{I^i}(\boldsymbol{a}_{I_i})\|_2^2 \\ &\quad \cdot \underbrace{\|\boldsymbol{a}_{I_{\ell+1}} - \mathrm{NN}_{I^\ell}(\boldsymbol{a}_{I_{\ell+1}})\|_2^2 \cdot \|\boldsymbol{a}_{I_\ell} - \mathrm{NN}_{I^\ell \cup \{I_{\ell+1}\}}(\boldsymbol{a}_{I_\ell})\|_2^2}_{\neq \|\boldsymbol{a}_{I_\ell} - \mathrm{NN}_{I^\ell}(\boldsymbol{a}_{I_\ell})\|_2^2 \cdot \|\boldsymbol{a}_{I_{\ell+1}} - \mathrm{NN}_{I^{\ell+1}}(\boldsymbol{a}_{I_{\ell+1}})\|_2^2} \\ &\quad \cdot \prod_{i=\ell+2}^{k} \|\boldsymbol{a}_{I_i} - \mathrm{NN}_{I^i}(\boldsymbol{a}_{I_i})\|_2^2 \end{aligned} \qquad (40)$$

($I \cup \{j\}$ indicates that element j is put at the end of the sequence). We want to quantify the maximal increase in $N(s_\ell(I))$ compared to $N(I)$. The following Lemma shows that the maximal increase ratio is actually a constant, and thus does not depend on the data.

Lemma 7 *The following holds true:*

$$\begin{aligned} N(s_1(I)) &= N(I) \ , & (41) \\ N(s_\ell(I)) &\leq (1+\eta)^2 N(I) \ , \forall 2 \leq \ell \leq k-1 \ . & (42) \end{aligned}$$

Here, $0 \leq \eta \leq 3$ *is a constant.*

The proof stems directly from the following Lemma.

Lemma 8 *For any non-empty* $\mathcal{N} \subseteq \mathcal{A}$ *and* $x \in \Omega$*, let* $\mathrm{NN}_{\mathcal{N}}(x)$ *denote the nearest neighbor of* x *in* $\mathcal{N}$*. There exists a constant* $0 \leq \eta \leq 3$ *such that for any* $\boldsymbol{a}_i, \boldsymbol{a}_j \in \mathcal{A}$ *and any nonempty subset* $\mathcal{N} \subseteq \mathcal{A} \backslash \{\boldsymbol{a}_i, \boldsymbol{a}_j\}$*,*

$$\frac{\|\boldsymbol{a}_i - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2}{\|\boldsymbol{a}_i - \mathrm{NN}_{\mathcal{N} \cup \{\boldsymbol{a}_j\}}(\boldsymbol{a}_i)\|_2} \leq (1+\eta) \cdot \frac{\|\boldsymbol{a}_j - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_j)\|_2}{\|\boldsymbol{a}_j - \mathrm{NN}_{\mathcal{N} \cup \{\boldsymbol{a}_i\}}(\boldsymbol{a}_j)\|_2} \ . \qquad (43)$$

Proof Since $\|\boldsymbol{a}_j - \mathrm{NN}_{\mathcal{N} \cup \{\boldsymbol{a}_i\}}(\boldsymbol{a}_j)\|_2 \leq \|\boldsymbol{a}_j - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_j)\|_2$, the proof is true for $\eta = 0$ when $\mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_i) = \mathrm{NN}_{\mathcal{N} \cup \{\boldsymbol{a}_j\}}(\boldsymbol{a}_i)$. So suppose that $\mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_i) \neq \mathrm{NN}_{\mathcal{N} \cup \{\boldsymbol{a}_j\}}(\boldsymbol{a}_i)$, implying $\mathrm{NN}_{\mathcal{N} \cup \{\boldsymbol{a}_j\}}(\boldsymbol{a}_i) = \boldsymbol{a}_j$. We distinguish two cases.

Case 1/2, if $\mathrm{NN}_{\mathcal{N} \cup \{\boldsymbol{a}_i\}}(\boldsymbol{a}_j) = \boldsymbol{a}_i$, then we are reduced to showing that $\|\boldsymbol{a}_i - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2 \leq (1+\eta)\|\boldsymbol{a}_j - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_j)\|_2$ under the conditions (C) that $\mathcal{N} \cap B(\boldsymbol{a}_i, \|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2) = \emptyset$ and $\mathcal{N} \cap B(\boldsymbol{a}_j, \|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2) = \emptyset$. Here, $B(\boldsymbol{a}, r)$ denotes the open ball of center $\boldsymbol{a}$ and radius R. The triangle inequality and conditions (C) bring

$$\begin{aligned} \|\boldsymbol{a}_i - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2 &\leq \|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2 + \|\boldsymbol{a}_j - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2 \\ &\leq \|\boldsymbol{a}_j - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_j)\|_2 + \|\boldsymbol{a}_j - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2 \ . \end{aligned} \qquad (44)$$

If $\mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_i) = \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_j)$ then the inequality holds for $\eta = 1$. Otherwise, suppose that $\|\boldsymbol{a}_j - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2 > 3\|\boldsymbol{a}_j - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_j)\|_2$. The triangle inequality yields again $\|\boldsymbol{a}_j - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2 \leq \|\boldsymbol{a}_j - \boldsymbol{a}_i\|_2 + \|\boldsymbol{a}_i - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2$, and so we have the inequality:

$$3\|\boldsymbol{a}_j - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_j)\|_2 < \|\boldsymbol{a}_j - \boldsymbol{a}_i\|_2 + \|\boldsymbol{a}_i - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2 \ , \qquad (45)$$

and since (C) holds, $\|\boldsymbol{a}_j - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_j)\|_2 \geq \|\boldsymbol{a}_j - \boldsymbol{a}_i\|_2$ which implies

$$\|\boldsymbol{a}_j - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_j)\|_2 < \frac{1}{2} \cdot \|\boldsymbol{a}_i - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2 \ . \tag{46}$$

On the other hand, the triangle inequality brings again

$$\begin{aligned} \|\boldsymbol{a}_i - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_j)\|_2 &\leq \|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2 + \|\boldsymbol{a}_j - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_j)\|_2 \\ &\leq 2 \cdot \|\boldsymbol{a}_j - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_j)\|_2 \qquad (47) \\ &< 2 \cdot \frac{1}{2} \cdot \|\boldsymbol{a}_i - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2 = \|\boldsymbol{a}_i - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2 \ , \qquad (48) \end{aligned}$$

a contradiction since $\|\boldsymbol{a}_i - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2 \leq \|\boldsymbol{a}_i - \boldsymbol{a}_l\|_2, \forall \boldsymbol{a}_l \in \mathcal{N}$ by definition. Ineq. (47) uses (C) and ineq. (48) uses ineq. (46). Hence, if $\textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_i) \neq \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_j)$ then since $\|\boldsymbol{a}_j - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2 \leq 3\|\boldsymbol{a}_j - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_j)\|_2$, ineq. (44) brings $\|\boldsymbol{a}_i - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2 \leq 4 \cdot \|\boldsymbol{a}_j - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_j)\|_2$, and the inequality holds for $\eta = 3$.

Case 2/2, if $\textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{a}_i\}}(\boldsymbol{a}_j) \neq \boldsymbol{a}_i$, then it implies $\textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{a}_i\}}(\boldsymbol{a}_j) = \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_j)$ and so

$$\exists \boldsymbol{a}_* \in \mathcal{N} : \|\boldsymbol{a}_j - \boldsymbol{a}_*\|_2 \ \ \leq \|\boldsymbol{a}_j - \boldsymbol{a}_i\|_2 \ . \tag{49}$$

Ineq. (43) reduces to proving

$$\|\boldsymbol{a}_i - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2 \ \ \leq \ \ (1+\eta) \cdot \|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2 \ , \tag{50}$$

but $\|\boldsymbol{a}_i - \boldsymbol{a}_*\|_2 \leq \|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2 + \|\boldsymbol{a}_j - \boldsymbol{a}_*\|_2 \leq 2\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2$, and since $\boldsymbol{a}_* \in \mathcal{N}$, $\|\boldsymbol{a}_i - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a}_i)\|_2 \leq \|\boldsymbol{a}_i - \boldsymbol{a}_*\|_2 \leq 2\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2$, and (50) is proved for $\eta = 1$. This achieves the proof of Lemma 8. ■

Let I be any sequence not containing the index of $\boldsymbol{a}'_n$, and let $I(i)$ denote the sequence in which we replace $\boldsymbol{a}_{I_i}$ by the index of $\boldsymbol{a}'_n$. The sequence of swaps

$$I(k) \ \ = \ \ (s_{k-1} \circ ... \circ s_{i+1} \circ s_i)(I(i)) \tag{51}$$

produces a sequence $I(k)$ in which all elements different from $\boldsymbol{a}'_n$ are in the same relative order as they are in I with respect to each other, and $\boldsymbol{a}'_n$ is pushed to the end of the sequence in k^{th} rank. We also have

$$N(I(i)) \ \ \leq \ \ (1+\eta)^{2(k-i)} N(I(k)) \ . \tag{52}$$

All the properties we need on $N(.)$ are now established. We turn to the analysis of $M(I^i|\mathcal{A})$.

Lemma 9 *For any $\delta_s > 0$ such that $\mathcal{A}$ is δ_s-monotonic, the following holds. For any $\mathcal{N} \subseteq \mathcal{A}$ with $|\mathcal{N}| \in \{1, 2, ..., k-1\}$, $\forall \boldsymbol{x}, \boldsymbol{x}' \in \Omega$, we have:*

$$\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}\}}(\boldsymbol{a})\|_2^2 \ \ \leq \ \ (1+\delta_s) \cdot \sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}'\}}(\boldsymbol{a})\|_2^2 \ . \tag{53}$$

Proof Since adding a point to $\mathcal{N}$ cannot increase the potential $\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}\}}(\boldsymbol{a})\|_2^2$, it comes

$$\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}\}}(\boldsymbol{a})\|_2^2 \ \ \leq \ \ \sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a})\|_2^2 \ , \forall \boldsymbol{x} \in \Omega \ . \tag{54}$$

Consider any $\boldsymbol{x}' \in \Omega$ such that $\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}'\}}(\boldsymbol{a})\|_2^2 = \sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a})\|_2^2$, *i.e.*, all points of $\mathcal{A}$ are closer to a point in $\mathcal{N}$ than they are from $\boldsymbol{x}'$. In this case, we obtain from ineq. (54),

$$\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}\}}(\boldsymbol{a})\|_2^2 \;\; \leq \;\; \sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}'\}}(\boldsymbol{a})\|_2^2 \ , \tag{55}$$

and since $\delta_s > 0$, the statement of the Lemma holds.

More interesting is the case where $\boldsymbol{x}' \in \Omega$ is such that $\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}'\}}(\boldsymbol{a})\|_2^2 < \sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a})\|_2^2$, implying $\boldsymbol{x}' \not\in \mathcal{N}$. In this case, let $A \doteq \{\boldsymbol{a} \in \mathcal{A} : \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}'\}}(\boldsymbol{a}) = \boldsymbol{x}'\}$, which is then non-empty. Let us denote for short $\boldsymbol{c}(A) \doteq (1/|A|) \cdot \sum_{\boldsymbol{a}\in A} \boldsymbol{a}$. Since $\boldsymbol{x}' \not\in \mathcal{N}$, $A \cap \mathcal{N} = \emptyset$, and since $\mathcal{A}$ is δ_s-monotonic, then it comes from ineq. (54)

$$\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}\}}(\boldsymbol{a})\|_2^2 \;\; \leq \;\; (1+\delta_s) \cdot \sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{c}(A)\}}(\boldsymbol{a})\|_2^2 \ . \tag{56}$$

We have:

$$\begin{aligned}
\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{c}(A)\}}(\boldsymbol{a})\|_2^2 \;\; &= \;\; \sum_{\boldsymbol{a}\in\mathcal{A}\backslash A} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{c}(A)\}}(\boldsymbol{a})\|_2^2 + \sum_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{c}(A)\}}(\boldsymbol{a})\|_2^2 \\
&\leq \;\; \sum_{\boldsymbol{a}\in\mathcal{A}\backslash A} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{c}(A)\}}(\boldsymbol{a})\|_2^2 + \sum_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \boldsymbol{c}(A)\|_2^2 \\
&\leq \;\; \sum_{\boldsymbol{a}\in\mathcal{A}\backslash A} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{c}(A)\}}(\boldsymbol{a})\|_2^2 + \sum_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \boldsymbol{x}'\|_2^2 \ .
\end{aligned} \tag{57}$$

Eq. (57) holds because the arithmetic average is the population minimizer of L_2^2. Because of the definition of A,

$$\begin{aligned}
\sum_{\boldsymbol{a}\in\mathcal{A}\backslash A} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{c}(A)\}}(\boldsymbol{a})\|_2^2 \;\; &\leq \;\; \sum_{\boldsymbol{a}\in\mathcal{A}\backslash A} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a})\|_2^2 \\
&= \sum_{\boldsymbol{a}\in\mathcal{A}\backslash A} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}'\}}(\boldsymbol{a})\|_2^2 \ ,
\end{aligned} \tag{58}$$

and, still because of the definition of A,

$$\sum_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \boldsymbol{x}'\|_2^2 \;\; = \;\; \sum_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}'\}}(\boldsymbol{a})\|_2^2 \ , \tag{59}$$

so we get from (58) and (59) $\sum_{\boldsymbol{a}\in\mathcal{A}\backslash A} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{c}(A)\}}(\boldsymbol{a})\|_2^2 + \sum_{\boldsymbol{a}\in A} \|\boldsymbol{a} - \boldsymbol{x}'\|_2^2 \leq \sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}'\}}(\boldsymbol{a})\|_2^2$, and finally from ineq. (57),

$$\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{c}(A)\}}(\boldsymbol{a})\|_2^2 \;\; \leq \;\; \sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}'\}}(\boldsymbol{a})\|_2^2 \ , \tag{60}$$

which, using ineq. (56), completes the proof of Lemma 9. ■

Lemma 10 *The following holds true, for any $i > 1$, any $\mathcal{A}' \approx \mathcal{A}$, any $\delta_w, \delta_s > 0$:*

$$\begin{aligned}
\mathcal{A} \textit{ is } \delta_w\textit{-spread} \quad &\Rightarrow \quad (n \notin I^i \Rightarrow M(I^i|\mathcal{A}) \leq (1+\delta_w) \cdot M(I^i|\mathcal{A}')) \ , && (61)\\
\mathcal{A} \textit{ is } \delta_s\textit{-monotonic} \quad &\Rightarrow \quad (n \in I^i \Rightarrow M(I^i|\mathcal{A}) \leq (1+\delta_s) \cdot M(I^i|\mathcal{A}')) \ . && (62)
\end{aligned}$$

Proof Suppose first that $n \notin I^i$. In this case, since $\mathcal{A}$ is δ_w-spread,

$$\begin{aligned}
M(I^i|\mathcal{A}) &= \sum_{j=1}^{n} \|\boldsymbol{a}_j - \textsc{nn}_{I^i}(\boldsymbol{a}_j)\|_2^2\\
&= \sum_{j=1}^{n-1} \|\boldsymbol{a}_j - \textsc{nn}_{I^i}(\boldsymbol{a}_j)\|_2^2 + \|\boldsymbol{a}_n - \textsc{nn}_{I^i}(\boldsymbol{a}_j)\|_2^2\\
&\leq \sum_{j=1}^{n-1} \|\boldsymbol{a}_j - \textsc{nn}_{I^i}(\boldsymbol{a}_j)\|_2^2 + R^2\\
&\leq (1+\delta_w) \cdot \sum_{j=1}^{n-1} \|\boldsymbol{a}_j - \textsc{nn}_{I^i}(\boldsymbol{a}_j)\|_2^2 && (63)\\
&\leq (1+\delta_w) \cdot \left(\sum_{j=1}^{n-1} \|\boldsymbol{a}_j - \textsc{nn}_{I^i}(\boldsymbol{a}_j)\|_2^2 + \|\boldsymbol{a}'_n - \textsc{nn}_{I^i}(\boldsymbol{a}'_n)\|_2^2\right)\\
&\quad = (1+\delta_w) \cdot M(I^i|\mathcal{A}') \ , && (64)
\end{aligned}$$

as indeed computing the nearest neighbors do not involve the n^{th} element of the sets, *i.e.* $\boldsymbol{a}_n$ or $\boldsymbol{a}'_n$. We have used in ineq. (63) the fact that $\mathcal{A}$ is δ_w-spread.

When $n \in I^i$, eq. (62) is an immediate consequence of Lemma 9 in which the distinct elements of $\mathcal{A}$ and $\mathcal{A}'$ play the role of $\boldsymbol{x}$ and $\boldsymbol{x}'$. ■

Lemma 11 *For any $\delta_w > 0$, if $\mathcal{A}$ is δ_w-spread, then for any $\mathcal{N} \subseteq \mathcal{A}$ with $|\mathcal{N}| = k-1$, $\forall \boldsymbol{x} \in \Omega$, it holds that $\|\boldsymbol{x} - \textsc{nn}_{\mathcal{N}}(\boldsymbol{x})\|_2^2 \leq \delta_w \sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a})\|_2^2$.*

Proof Follows directly from the fact that $\|\boldsymbol{x} - \textsc{nn}_{\mathcal{N}}(\boldsymbol{x})\|_2^2 \leq R^2$ by assumption. ■

Letting $I(k)$ denote a sequence containing element n pushed to the end of the sequence, we get:

$$\begin{aligned}
&\sum_{\boldsymbol{\sigma}\in S_k} \sum_{I \in Seq^+(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A}')\\
&\quad = \sum_{\boldsymbol{\sigma}\in S_k} \sum_{I \in Seq^+(n:k)} \frac{N(I)}{\prod_{i=1}^{k} M(I^i|\mathcal{A}')} \cdot p_{\boldsymbol{a'}_n}(\boldsymbol{c}_{\sigma(i)}) \cdot \prod_{i=1:I_i\neq n}^{k} p_{\boldsymbol{a}_{I_i}}(\boldsymbol{c}_{\sigma(i)})\\
&\quad \leq (1+\eta)^{2(k-2)}\\
&\qquad \cdot \sum_{\boldsymbol{\sigma}\in S_k} \sum_{I \in Seq^+(n:k)} \frac{N(I(k))}{\prod_{i=1}^{k} M(I^i|\mathcal{A}')} \cdot p_{\boldsymbol{a'}_n}(\boldsymbol{c}_{\sigma(i)}) \cdot \prod_{i=1:I_i\neq n}^{k} p_{\boldsymbol{a}_{I_i}}(\boldsymbol{c}_{\sigma(i)}) \ . && (65)
\end{aligned}$$

Now, take any element $I \in Seq_{+}(n:k)$ with $\boldsymbol{a}'_n$ in position k, and change $\boldsymbol{a}'_n$ by some $\boldsymbol{a} \in \mathcal{A}$. Any of these changes generates a different element $I' \in Seq^{-}(n:k)$, and so using Lemma 11 and the following two facts:

- the fact that

$$p_{\boldsymbol{a}'_n}(\boldsymbol{c}_{\sigma(i)}) \leq \varrho(R) \cdot p_{\boldsymbol{a}}(\boldsymbol{c}_{\sigma(i)}) \ , \tag{66}$$

  for any $\boldsymbol{a} \in \mathcal{A}$,

- the fact that, if $\mathcal{A}$ is δ_s-monotonic,

$$M(I^i_{\boldsymbol{a}}|\mathcal{A}) \leq (1+\delta_s) \cdot M(I^i|\mathcal{A}) \ , \tag{67}$$

  for any $\boldsymbol{a} \in \mathcal{A}$ not already in the sequence, where $I_{\boldsymbol{a}}$ denotes the sequence I in which $\boldsymbol{a}'_n$ has been replaced by $\boldsymbol{a}$,

we get from ineq. (65),

$$\begin{aligned} &\sum_{\boldsymbol{\sigma} \in S_k} \sum_{I \in Seq^{+}(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A}') \\ &\quad \leq (1+\eta)^{2(k-2)} \cdot (1+\delta_s)^{k-1} \cdot \delta_w \\ &\qquad \cdot \varrho(R) \cdot \sum_{\boldsymbol{\sigma} \in S_k} \sum_{I \in Seq^{-}(n:k)} \frac{N(I)}{\prod_{i=1}^{k} M(I^i|\mathcal{A})} \cdot \prod_{i=1}^{k} p_{\boldsymbol{a}_{I_i}}(\boldsymbol{c}_{\sigma(i)}) \ . \end{aligned} \tag{68}$$

Lemma 12 *For any $\delta_w, \delta_s > 0$ such that $\mathcal{A}$ is δ_w-spread and δ_s-monotonic, for any $\mathcal{A}' \approx \mathcal{A}$, we have:*

$$\frac{\mathbb{P}[\mathcal{C}|\mathcal{A}']}{\mathbb{P}[\mathcal{C}|\mathcal{A}]} \leq (1+\delta_w)^{k-1} \cdot \left(1 + \delta_w \cdot \left(\frac{1+\delta_s}{1+\delta_w}\right)^{k-1} \cdot (1+\eta)^{2(k-2)} \cdot \varrho(R)\right) \ . \tag{69}$$

Proof We get from the fact that $\mathcal{A}$ is δ_w-spread,

$$\sum_{\boldsymbol{\sigma} \in S_k} \sum_{I \in Seq_{-}(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A}') \leq (1+\delta_w)^{k-1} \cdot \sum_{\boldsymbol{\sigma} \in S_k} \sum_{I \in Seq_{-}(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A}) \ , \tag{70}$$

and furthermore ineq. (68) yields:

$$\begin{aligned}
\frac{\mathbb{P}[\mathcal{C}|\mathcal{A}']}{\mathbb{P}[\mathcal{C}|\mathcal{A}]} &= \frac{\sum_{\boldsymbol{\sigma}\in S_k}\sum_{I\in Seq(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A}')}{\sum_{\boldsymbol{\sigma}\in S_k}\sum_{I\in Seq(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A})} \\
&\leq \frac{\left(\begin{array}{c} (1+\delta_w)^{k-1}\cdot \sum_{\boldsymbol{\sigma}\in S_k}\sum_{I\in Seq_-(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A}) \\ + \\ \sum_{\boldsymbol{\sigma}\in S_k}\sum_{I\in Seq_+(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A}') \end{array}\right)}{\sum_{\boldsymbol{\sigma}\in S_k}\sum_{I\in Seq(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A})} \\
&\leq (1+\delta_w)^{k-1} \\
&\quad \cdot \frac{\left(\begin{array}{c} \sum_{\boldsymbol{\sigma}\in S_k}\sum_{I\in Seq_-(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A}) \\ + \\ \delta_w \cdot \left(\frac{1+\delta_s}{1+\delta_w}\right)^{k-1} \cdot (1+\eta)^{2(k-2)} \cdot \varrho(R) \cdot \sum_{\boldsymbol{\sigma}\in S_k}\sum_{I\in Seq_-(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A}') \end{array}\right)}{\sum_{\boldsymbol{\sigma}\in S_k}\sum_{I\in Seq(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A})} \\
&= (1+\delta_w)^{k-1} \cdot \left(1 + \delta_w \cdot \left(\frac{1+\delta_s}{1+\delta_w}\right)^{k-1} \cdot (1+\eta)^{2(k-2)} \cdot \varrho(R)\right) \\
&\quad \cdot \underbrace{\frac{\sum_{\boldsymbol{\sigma}\in S_k}\sum_{I\in Seq_-(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A})}{\sum_{\boldsymbol{\sigma}\in S_k}\sum_{I\in Seq(n:k)} p(\boldsymbol{\sigma}, I, \mathcal{C}|\mathcal{A})}}_{\leq 1} .
\end{aligned}$$

This ends the proof of Lemma 12. ■

Since

$$\begin{aligned}
& (1+\delta_w)^{k-1} \cdot \left(1 + \delta_w \cdot \left(\frac{1+\delta_s}{1+\delta_w}\right)^{k-1} \cdot (1+\eta)^{2(k-2)} \cdot \varrho(R)\right) \\
&= (1+\delta_w)^{k-1} + (1+\eta)^{2(k-2)} \cdot \delta_w \cdot (1+\delta_s)^{k-1} \cdot \varrho(R) ,
\end{aligned}$$

and $\eta \leq 3$ from Lemma 7, we get Theorem 9 with

$$f(k) \doteq 4^{2k-4} . \tag{71}$$

2.6 Proof of Theorem 10

Assume that density $\mathcal{D}$ contains a L_2 ball $\mathscr{B}_2(\mathbf{0}, R)$ of radius R, centered without loss of generality in $\mathbf{0}$. Fix $0 < \kappa < m - 1$. For any $\alpha \in (0, 1)$ and $\mathcal{N} \subseteq \mathcal{A}$ with $|\mathcal{N}| \in \{1, 2, ..., \kappa\} \doteq [\kappa]_*$, let $\mathcal{N} \oplus \alpha \doteq \cup_{\boldsymbol{x}\in\mathcal{N}} \mathscr{B}_2(\boldsymbol{x}, \alpha \cdot R)$ be the union of all small balls centered around each element of $\mathcal{N}$, each of radius $\alpha \cdot R$. An important quantity is

$$q_* \doteq \min_{\mathcal{N}\subseteq\mathcal{A}, |\mathcal{N}|\in[\kappa]_*} \frac{\mu(\mathscr{B}_2(\mathbf{0}, R)\backslash\mathcal{N} \oplus \alpha)}{\mu(\mathscr{B}_2(\mathbf{0}, R))} \tag{72}$$

the minimal mass of $\mathscr{B}_2(\mathbf{0}, R)\backslash\mathcal{N} \oplus \alpha$ relatively to $\mathscr{B}_2(\mathbf{0}, R)$ as measured using $\mathcal{D}$. As depicted in Figure 4, q_* is a minimal value of the probability to escape the neighborhoods of $\mathcal{N} \oplus \alpha$ when

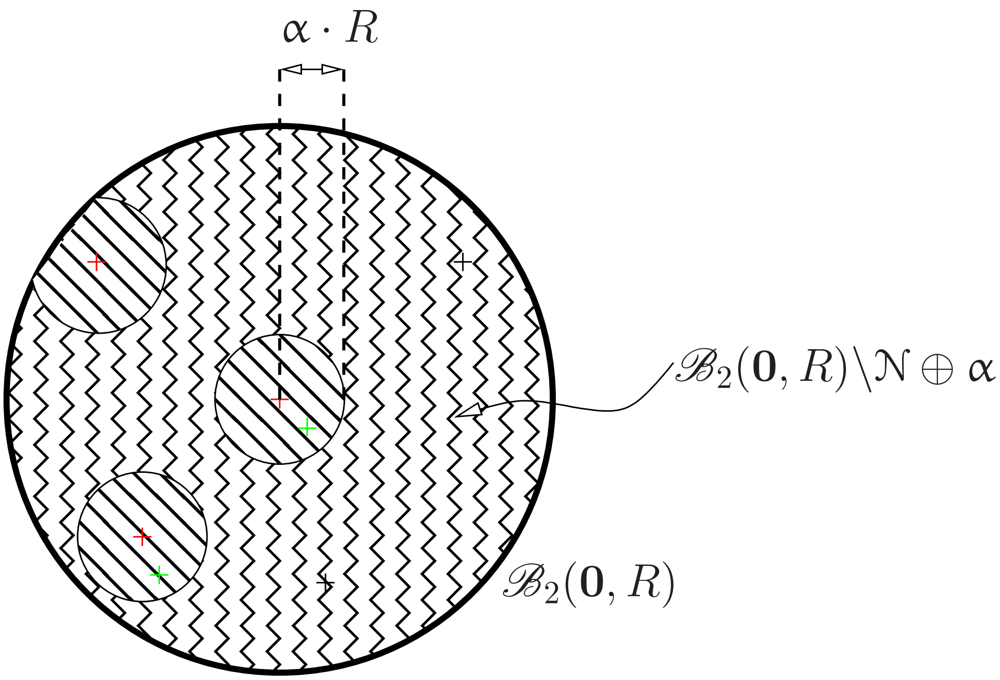

$\alpha \cdot R$
$\mathscr{B}_2(\mathbf{0}, R)\backslash\mathcal{N} \oplus \alpha$
$\mathscr{B}_2(\mathbf{0}, R)$

Figure 4: q_* in eq. (72) measures the probability the a point drawn in $\mathscr{B}_2(\mathbf{0}, R)$ escapes the neighborhoods of $\mathcal{N} \oplus \alpha$. In this example, two points in black escape the neighborhoods (defined by three points in red), while two in green do not.

sampling points according to $\mathcal{D}$ in ball $\mathscr{B}_2(\mathbf{0}, R)$. If, for some α that shall depend upon the dimension d and κ, q_* is large enough, then the spread of points drawn shall guarantee "small" values for δ_w and δ_s.

This is formalized in the following Theorem, which assumes $\epsilon_m = \epsilon_M = 1$, *i.e.* the ball has uniform density. Theorem 10 is a direct consequence of this Theorem.

Theorem 13 *Suppose $\mathcal{A} \subset \mathscr{B}_2(\mathbf{0}, R)$. For any $\delta \in (0, 1)$, if*

$$m \geq 3\left(\frac{\kappa}{q_*\delta^2}\right)^2 , \tag{73}$$

then there is probability $\geq 1 - \delta$ over its sampling that $\mathcal{A}$ is δ_w-spread and δ_s-monotonic for the following values of δ_w, δ_s:

$$\delta_w = \frac{1}{q_*(1-\delta)(m-\kappa-1)\alpha^2} , \tag{74}$$

$$\delta_s = \frac{m}{m-\kappa} \cdot \left(\frac{2}{\min\left\{\frac{1}{4}, q_*(1-\delta)\right\} \cdot \alpha}\right)^2 - 1 . \tag{75}$$

Proof We first prove the following Lemma.

Lemma 14 *Suppose $\mathcal{A} \subset \mathscr{B}_2(\mathbf{0}, R)$. Let q_* be defined as in eq. (72). Then for any $\delta \in (0,1)$, if m meets ineq. (73), then there is probability $\geq 1 - \delta$ that*

$$|(\mathscr{B}_2(\mathbf{0}, R)\backslash\mathcal{N} \oplus \alpha) \cap (\mathcal{A}\backslash\mathcal{N})| \geq q_*(1-\delta)(m-\kappa) \ , \forall \mathcal{N} \subseteq \mathcal{A}, |\mathcal{N}| \in [\kappa]_* \ . \tag{76}$$

Proof Since we assume $\mathcal{A} \subset \mathscr{B}_2(\mathbf{0}, R)$, Chernoff bounds imply that for any *fixed* $\mathcal{N} \subseteq \mathcal{A}$ with $|\mathcal{N}| \in [\kappa]_*$,

$$\mathbb{P}_{\mathcal{D}}\left[\frac{|(\mathscr{B}_2(\mathbf{0}, R)\backslash\mathcal{N} \oplus \alpha) \cap (\mathcal{A}\backslash\mathcal{N})|}{|\mathcal{A}\backslash\mathcal{N}|} \leq q_*(1-\delta)\right] \leq \exp\left(-\delta^2 q_* \left|\mathcal{A}\backslash\mathcal{N}\right|/2\right) \ . \tag{77}$$

Now, remark that

$$\sum_{j=1}^{\kappa} \binom{m}{j} \leq m^{\kappa} \ , \forall m, \kappa \geq 1 \ . \tag{78}$$

This can be proven by induction, m being fixed: it trivially holds for $\kappa = 1$ and $\kappa = 2$, and furthermore

$$\begin{aligned}\sum_{j=1}^{\kappa} \binom{m}{j} &= \sum_{j=1}^{\kappa-1} \binom{m}{j} + \binom{m}{\kappa} \\ &\leq m^{\kappa-1} + \frac{m!}{(m-\kappa)!\kappa!} \ ,\end{aligned} \tag{79}$$

by induction at rank $\kappa - 1$. To prove that the right-hand side of (79) is no more than m^{κ}, we just have to remark that

$$\begin{aligned}\frac{m!}{(m-\kappa)!\kappa! m^{\kappa-1}} &< \frac{m}{\kappa!} \\ &\leq m-1 \ ,\end{aligned} \tag{80}$$

as long as $\kappa > 1$ and $m > 1$. So, the property at rank $\kappa - 1$ for $\kappa > 1$ implies property at rank κ, which concludes the induction.

So, we have at most m^{κ} choices for $\mathcal{N}$, so relaxing the choice of $\mathcal{N}$, we get

$$\begin{aligned}&\mathbb{P}_{\mathcal{D}}\left[\exists \mathcal{N} \subseteq \mathcal{A}, |\mathcal{N}| = \kappa : \frac{|(\mathscr{B}_2(\mathbf{0}, R)\backslash\mathcal{N} \oplus \alpha) \cap \mathcal{A}_{\mathcal{N}}|}{|\mathcal{A}_{\mathcal{N}}|} \leq q_*(1-\delta)\right] \\ &\leq m^{\kappa} \exp\left(-\frac{\delta^2 q_*(m-\kappa)}{2}\right) \ .\end{aligned} \tag{81}$$

We want to compute the minimal m such that the right-hand side is no more than δ, this being equivalent to

$$\delta^2 q_* m \geq 2\log\left(\frac{m^{\kappa}}{\delta}\right) + \kappa\delta^2 q_* \ ,$$

which, since $\delta \in (0,1)$, is ensured if

$$\delta^2 q_* m \geq 2\kappa\log\left(\frac{m}{\delta}\right) + \kappa\delta^2 q_* \ . \tag{82}$$

Suppose

$$m \;=\; 3\left(\frac{\kappa}{q_*\delta^2}\right)^2 \;.$$

Since we trivially have $\kappa^2/(q_*\delta^2)^2 \geq \kappa\delta^2 q_*$ ($\kappa \geq 1, q_* \in (0,1), \delta \in (0,1)$), it is sufficient to prove:

$$\frac{2\kappa}{q_*\delta^2} \;\geq\; 2\log 3 + 2\log\left(\frac{\kappa^2}{q_*^2\delta^5}\right) \;, \qquad (83)$$

which, again observing that $\delta \in (0,1)$, holds if we can prove

$$\frac{\kappa}{q_*\delta^2} \;\geq\; \log 2 + \frac{3}{2}\cdot\log\left(\frac{\kappa}{q_*\delta^2}\right) \;, \qquad (84)$$

which is equivalent to showing $x \geq (3/2)\log x + \log 2$ for $x \geq 1$, which indeed holds (end of the proof of Lemma 14). ■

The consequence of Lemma 14 is the following: if $\mathcal{A} \subset \mathscr{B}_2(\mathbf{0}, R)$ and m satisfies (73), then for any $\mathcal{N} \subseteq \mathcal{A}$ with $|\mathcal{N}| = k - 1$, and any $\mathcal{B} \subseteq \mathcal{A}$ with $|\mathcal{B}| = |\mathcal{A}| - 1$,

$$\sum_{\boldsymbol{a}\in\mathcal{B}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}}(\boldsymbol{a})\|_2^2 \;\geq\; q_*(1-\delta)(m-\kappa-1)\alpha^2 \cdot R^2 \;, \qquad (85)$$

and so from Definition 7 (main file) $\mathcal{A}$ is δ_s-spread for:

$$\delta_w \;=\; \frac{1}{q_*(1-\delta)(m-\kappa-1)\alpha^2} \;. \qquad (86)$$

Now, suppose we add a single point $\boldsymbol{x}_*$ in $\mathcal{N}$. If, for some fixed $\alpha_* \in (0, \alpha/2]$,

$$\boldsymbol{x}_* \;\not\in\; \boldsymbol{a} \oplus \alpha_* \;, \forall \boldsymbol{a} \in \mathcal{A} \;, \qquad (87)$$

then because of (85),

$$\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}_*\}}(\boldsymbol{a})\|_2^2 \;\geq\; (m-\kappa)\cdot\min\left\{\alpha_*^2, q_*(1-\delta)\alpha^2\right\}\cdot R^2 \;. \qquad (88)$$

Otherwise, consider one $\boldsymbol{a}_*$ for which $\boldsymbol{x}_* \in \boldsymbol{a}_* \oplus \alpha_*$. If we replace $\boldsymbol{a}_*$ by $\boldsymbol{x}_*$ in all $\mathcal{N}$ in which $\boldsymbol{a}_*$ belongs to in Lemma 14, then because $\boldsymbol{x}_* \oplus \alpha_* \subset \boldsymbol{a}_* \oplus \alpha$, it comes from Lemma 14:

$$\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{x}_*\}}(\boldsymbol{a})\|_2^2 \;\geq\; \frac{1}{4}\cdot(m-\kappa)\cdot q_*(1-\delta)\alpha^2\cdot R^2 \;. \qquad (89)$$

We thus get in all cases

$$\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \textsc{nn}_{\mathcal{N}\cup\{\boldsymbol{c}(A)\}}(\boldsymbol{a})\|_2^2 \;\geq\; \min\left\{\frac{\alpha^2}{4}, \alpha_*^2, q_*(1-\delta)\alpha^2\right\}(m-\kappa)\cdot q_*(1-\delta)\cdot R^2 \;, \qquad (90)$$

where $\boldsymbol{c}(A)$ is the arithmetic average computed according to the definition of δ_s-monotonicity, of any $A \subseteq \mathcal{A}\backslash\mathcal{N}$. Since $\mathcal{N} \subseteq \mathcal{A} \subset \mathscr{B}_2(\mathbf{0}, R)$, we have $\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a})\|_2^2 \leq 4mR^2$, and so

$$\sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \mathrm{NN}_{\mathcal{N}}(\boldsymbol{a})\|_2^2 \leq \frac{4m}{\min\left\{\frac{\alpha^2}{4}, \alpha_*^2, q_*(1-\delta)\alpha^2\right\}(m-\kappa)\cdot q_*(1-\delta)} \cdot \sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{a} - \mathrm{NN}_{\mathcal{N}\cup\{\boldsymbol{c}(A)\}}(\boldsymbol{a})\|_2^2 \tag{91}$$

implying from Definition 8 (main file) that δ_s-monotonicity holds with:

$$\delta_s = \frac{m}{m-\kappa} \cdot \frac{4}{\min\left\{\frac{\alpha^2}{4}, \alpha_*^2, q_*(1-\delta)\alpha^2\right\}\cdot q_*(1-\delta)} - 1 \ . \tag{92}$$

The statement of the Theorem follows with $\alpha_* = \alpha/2$ (end of the proof of Theorem 13). ■

We finish the proof of Theorem 10. We have

$$q_* \geq 1 - \kappa\alpha^d \ , \tag{93}$$

where the lowerbound corresponds to the case where all neighborhoods in $\mathcal{N} \oplus \alpha$ are distinct and included in $\mathscr{B}_2(\mathbf{0}, R)$. So we have, for any fixed choice of $\alpha \in (0, 1)$,

$$\delta_w \leq \frac{1}{\alpha^2 \cdot (1-\kappa\alpha^d)(1-\delta)(m-\kappa-1)} \ . \tag{94}$$

To minimize this upperbound, we pick α to maximize $\alpha^2 \cdot (1-\kappa\alpha^d)$ with $\alpha \in (0, 1)$, which is easily achieved picking

$$\alpha = \left(\frac{1}{\kappa(d+1)}\right)^{\frac{1}{d}} \ , \tag{95}$$

and yields

$$\begin{aligned}\delta_w &\leq \left(1+\frac{1}{d}\right)\cdot\frac{1}{(\kappa(d+1))^{\frac{2}{d}}(1-\delta)(m-\kappa-1)} \\ &\leq \left(1+\frac{1}{d}\right)\cdot\frac{1}{\kappa^{\frac{2}{d}}(1-\delta)(m-\kappa-1)} \ .\end{aligned} \tag{96}$$

But we have for this choice, $1 - \kappa\alpha^d = d/(d+1) \geq 1/2$, so as long as

$$\delta < 1/2 \ , \tag{97}$$

we shall have $q_*(1-\delta) > 1/4$ and so we shall have

$$\begin{aligned}\delta_s + 1 &= 64 \cdot \frac{m}{m-\kappa}\cdot\frac{1}{\alpha^2} \\ &\leq 64 \cdot \frac{m}{m-\kappa}\cdot\frac{1}{\kappa^{\frac{2}{d}}} \ .\end{aligned} \tag{98}$$

We now go back to ineq. (14) (main file), which reads:

$$\frac{\mathbb{P}[\mathcal{C}|\mathcal{A}']}{\mathbb{P}[\mathcal{C}|\mathcal{A}]} \leq \varrho_1 + \varrho_2 \ , \tag{99}$$

with

$$\begin{aligned}
\varrho_1 &\doteq (1+\delta_w)^{k-1} \ , & (100)\\
\varrho_2 &\doteq f(k)\cdot \delta_w \cdot (1+\delta_s)^{k-1}\cdot \varrho(R) \ . & (101)
\end{aligned}$$

We upperbound separately both terms.

Lemma 15 *Suppose ineqs (97) and (15) (main file) are met. Then*

$$\varrho_1 \ \leq \ 1 + \frac{4}{m^{\frac{1}{4}+\frac{1}{d+1}}} \ . \qquad (102)$$

Proof Since $d \geq 1$ and $\delta < 1/2$, we get from ineq. (96) (using $\kappa = k$)

$$\begin{aligned}
(1+\delta_w)^{k-1} &\leq \left(1+\left(1+\frac{1}{d}\right)\cdot\frac{1}{k^{\frac{2}{d}}(1-\delta)(m-k-1)}\right)^{k-1}\\
&\leq \left(1+\frac{2}{k^{\frac{2}{d}}(1-\delta)(m-k-1)}\right)^{k-1}\\
&\leq \left(1+\frac{4}{k^{\frac{2}{d}}(m-k-1)}\right)^{k-1} \ . \qquad (103)
\end{aligned}$$

Let $h(k)$ be the right-hand side of ineq. (103). $h(1)$ trivially meets ineq. (102). When $k \geq 2$, h decreases until $k = 2(m-1)/(d+2)$ and then increases. We thus just need to check ineq. (102) for $k = 2$ and $k = \sqrt{m}$ from ineq. (15) (main file). We get $h(2) = 1 + 4/(4^{1/d}(m-3))$. For ineq. (102) to be satisfied, we need to have $4^{1/d}(m-3) \geq m^{\frac{1}{4}+\frac{1}{d+1}}$, which holds if $m \geq 3 + m^{3/4}$ ($d \geq 1$), that is, $m \geq 8$. But since ineqs (97) and (15) (main file) are satisfied, we have $m \geq 16k^2/\delta^2 \geq 64k^2 \geq 64$, and so $h(2)$ satisfies ineq. (102).

There remains to check ineq. (102) for $k = \sqrt{m}$. We have

$$\begin{aligned}
h(\sqrt{m}) &= \left(1+\frac{4}{m^{\frac{1}{d}}(m-\sqrt{m}-1)}\right)^{\sqrt{m}-1}\\
&\leq \left(1+\frac{4}{m^{\frac{1}{d}}(m-\sqrt{m})}\right)^{\sqrt{m}}\\
&\leq \left(1+\frac{2}{\sqrt{m}\cdot m^{\frac{1}{4}+\frac{1}{d}}}\right)^{\sqrt{m}} \ , \qquad (104)
\end{aligned}$$

since any $m \geq 64$, we have $m - \sqrt{m} \geq 2m^{3/4}$. To conclude, ineq (104) yields

$$\begin{aligned}
h(\sqrt{m}) &\leq \left(1+\frac{2}{\sqrt{m}\cdot m^{\frac{1}{4}+\frac{1}{d}}}\right)^{\sqrt{m}}\\
&\leq \exp\left(\frac{2}{m^{\frac{1}{4}+\frac{1}{d}}}\right)\\
&\leq 1+\frac{4}{m^{\frac{1}{4}+\frac{1}{d}}} \ . \qquad (105)
\end{aligned}$$

The penultimate ineq. comes from $1 + x \leq \exp x$, and the last one comes from the fact that $\exp(2x) \leq 1 + 4x$ for $x \leq 1$. Since $m^{\frac{1}{4}+\frac{1}{d}} \geq m^{\frac{1}{4}+\frac{1}{d+1}}$, we obtain the statement of the Lemma for $h(\sqrt{m})$. This concludes the proof of Lemma 15. ■

Lemma 16 *Suppose ineqs (97) and (15) (main file) are met. Then*

$$\varrho_2 \leq \left(\frac{64}{k^{\frac{2}{d}}}\right)^k \cdot \frac{\varrho(2R)}{m} . \tag{106}$$

Proof We fix $\kappa = k$, use $f(k) = 4^{2k-4}$ (eq. 71), so we get

$$\begin{aligned}
\varrho_2 &= 4^{2k-2} \cdot \left(1 + \frac{1}{d}\right) \cdot \frac{1}{k^{\frac{2}{d}}(1-\delta)(m-k-1)} \cdot \left(64 \cdot \frac{m}{m-k} \cdot \frac{1}{k^{\frac{2}{d}}}\right)^{k-1} \cdot \varrho(2R) \\
&\leq 2 \cdot 64^{k-1} \cdot \left(1 + \frac{1}{d}\right) \cdot \frac{1}{k^{\frac{2k}{d}}(m-k-1)} \cdot \left(1 + \frac{k}{m-k}\right)^{k-1} \cdot \varrho(2R) \qquad (107) \\
&\leq \underbrace{4 \cdot \frac{1}{(m-k-1)} \cdot \left(1 + \frac{k}{m-k}\right)^{k-1}}_{\doteq \varrho_3} \cdot 64^{k-1} \cdot \frac{1}{k^{\frac{2k}{d}}} \cdot \varrho(2R) , \qquad (108)
\end{aligned}$$

using the fact that $\delta < 1/2$ and $d \geq 1$. Now, we also have

$$\begin{aligned}
\left(1 + \frac{k}{m-k}\right)^{k-1} &\leq \exp\left(\frac{k^2}{m-k}\right) \qquad (109) \\
&\leq e , \qquad (110)
\end{aligned}$$

as long as $k \leq (1/16) \cdot \sqrt{m}$, and furthermore, since $m \geq 64$ (see the proof of Lemma 15), we also have $1/(m-k-1) \leq 5/m$. We thus obtain

$$\begin{aligned}
\varrho_3 &\leq \frac{20e}{m} \\
&\leq \frac{64}{m} , \qquad (111)
\end{aligned}$$

which yields

$$\varrho_2 \leq \left(\frac{64}{k^{\frac{2}{d}}}\right)^k \cdot \frac{\varrho(2R)}{m} , \tag{112}$$

as claimed. ■

Putting altogether Lemmata 15 and 16, we get:

$$\frac{\mathbb{P}[\mathcal{C}|\mathcal{A}']}{\mathbb{P}[\mathcal{C}|\mathcal{A}]} \leq 1 + \frac{4}{m^{\frac{1}{4}+\frac{1}{d+1}}} + \left(\frac{64}{k^{\frac{2}{d}}}\right)^k \cdot \frac{\varrho(2R)}{m} , \tag{113}$$

as claimed. There remains to check that, with our choice of α, the constraint on m in (73) is satisfied if

$$m \geq \frac{12k^2}{\delta^4} \tag{114}$$

since $q_* \geq d/(d+1)$. We obtain the sufficient constraint on k:

$$k \leq \frac{\delta^2}{4} \cdot \sqrt{m} \ , \tag{115}$$

which proves Theorem 10 when $\epsilon_m = \epsilon_M = 1$.

When the density do not satisfy $\epsilon_m = \epsilon_M = 1$ we just have to remark that the lowerbound on q_* is now

$$q_* \leq \frac{\varepsilon_m}{\varepsilon_M} \cdot (1 - \kappa\alpha^d) \ . \tag{116}$$

Ineq. (96) becomes

$$\delta_w \leq \frac{\varepsilon_M}{\varepsilon_m} \cdot \left(1 + \frac{1}{d}\right) \cdot \frac{1}{\kappa^{\frac{2}{d}}(1-\delta)(m - \kappa - 1)} \ , \tag{117}$$

ineq. (98) becomes

$$\delta_s + 1 \leq \frac{\varepsilon_M}{\varepsilon_m} \cdot 64 \cdot \frac{m}{m - \kappa} \cdot \frac{1}{\kappa^{\frac{2}{d}}} \ . \tag{118}$$

So, the only difference with the $\epsilon_m = \epsilon_M = 1$ is the ratio $\varepsilon_M/\varepsilon_m$ (≥ 1) which multiplies all quantities of interest, and yields, in lieu of ineq. (113),

$$\frac{\mathbb{P}[\mathcal{C}|\mathcal{A}']}{\mathbb{P}[\mathcal{C}|\mathcal{A}]} \leq 1 + \left(\frac{\varepsilon_M}{\varepsilon_m}\right)^k \cdot \left(\frac{4}{m^{\frac{1}{4}+\frac{1}{d+1}}} + \left(\frac{64}{k^{\frac{2}{d}}}\right)^k \cdot \frac{\varrho(2R)}{m}\right) \ , \tag{119}$$

which is the statement of Theorem 10.

2.7 Proof of Theorem 12

When $p_{(\boldsymbol{\mu}_{\boldsymbol{a}}, \boldsymbol{\theta}_{\boldsymbol{a}})}$ is a product of Laplace distributions $Lap(b)$ (b being the *scale* parameter of the distribution [8]), condition in ineq. (13) (main file) becomes:

$$\begin{aligned}\frac{p_{(\boldsymbol{\mu}_{\boldsymbol{a}'}, \boldsymbol{\theta}_{\boldsymbol{a}'})}(\boldsymbol{x})}{p_{(\boldsymbol{\mu}_{\boldsymbol{a}}, \boldsymbol{\theta}_{\boldsymbol{a}})}(\boldsymbol{x})} &\leq \exp\left(\frac{\|\boldsymbol{a} - \boldsymbol{a}'\|_1}{b}\right) \\ &= \exp\left(\frac{\sqrt{2}\|\boldsymbol{a} - \boldsymbol{a}'\|_1}{\sigma_1}\right) \\ &\leq \exp\left(\frac{2\sqrt{2}R}{\sigma_1}\right) \ , \forall \boldsymbol{a}, \boldsymbol{a}' \in \mathcal{A}, \forall \boldsymbol{x} \in \Omega \ ,\end{aligned} \tag{120}$$

assuming $\mathcal{A} \subset \mathscr{B}_1(\mathbf{0}, R)$. Let us fix $\varrho(R) \doteq \exp\left(2\sqrt{2}R/\sigma_1\right)$. Since $\mathscr{B}_1(\mathbf{0}, R) \subset \mathscr{B}_2(\mathbf{0}, R)$ (the L_2 ball), we now want $(1+\delta_w)^{k-1} + f(k) \cdot \delta_w \cdot (1+\delta_s)^{k-1} \cdot \varrho(R) = \exp(\epsilon)$. Solving for σ_1 yields:

$$\sigma_1 = \frac{2\sqrt{2}R}{\log\left(\frac{\exp(\epsilon)-(1+\delta_w)^{k-1}}{f(k)\cdot\delta_w\cdot(1+\delta_s)^{k-1}}\right)} , \tag{121}$$

as claimed. The proof that k-variates++ meets ineq. (7) with

$$\Phi = \Phi_1 \doteq 8 \cdot \left(\phi_{\mathrm{opt}} + \frac{mR^2}{\tilde{\epsilon}^2}\right) \tag{122}$$

comes from a direct application of Theorem 2 (main file), with

$$\begin{aligned} \eta &= 0 , \\ \phi_{\mathrm{bias}} &= \phi_{\mathrm{opt}} , \\ \phi_{\mathrm{var}} &= m \cdot \left(\frac{2\sqrt{2}R}{\tilde{\epsilon}}\right)^2 . \end{aligned}$$

The statements for σ_2 and Φ_2 are direct applications of the Laplace mechanism properties [8, 9].

2.8 Extension to non-metric spaces

Since its inception, the k-means++ seeding technique has been successfully adapted to various distortion measures $D(\cdot\|\cdot)$ to handle non-Euclidean features [10, 11, 12]. Similarly, our extended seeding technique can be adapted to these scenarii: this boils down to putting the distortion as a free parameter of the algorithm, replacing $D_t(\boldsymbol{a})$ (eq. (1)) by $D_t(\boldsymbol{a}) \doteq \min_{\boldsymbol{a}' \in \mathcal{P}} D(\boldsymbol{a}\|\boldsymbol{a}')$. For example, by noticing that the squared Euclidean distance is merely an example of Bregman divergences (the well-known canonical divergences in information geometry of dually flat spaces), k-variates++ can be been extended to that family of dissimilarities [11]. But more interesting examples now appear, that build on constraints that distortions have to satisfy for certain problems, like the invariance to rotations of the coordinate space. This is all the more challenging in practice for clustering since sometimes **no-closed form** solution are available for some of these divergences. Because it bypasses the construction of the population minimisers, k-variates++ offers an elegant solution to the problem. Such hard distortions include the skew Jeffreys α-centroids [12]. This also include the recent class of total Bregman/Jensen divergences that are examples of conformal divergences [13, 12]. We give an example of the extension of k-variates++to the total Jensen divergence, to show that k-variates++ can approximate the optimal clustering even without closed form solutions for the population minimisers [13]. For any convex function $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ and $\alpha \in (0, 1)$, the skew Jensen divergence is

$$J_\alpha(\boldsymbol{a}, \boldsymbol{a}') \doteq \alpha\varphi(\boldsymbol{a}) + (1-\alpha)\varphi(\boldsymbol{a}') - \varphi(\alpha\boldsymbol{a} + (1-\alpha)\boldsymbol{a}') , \tag{123}$$

and the total Jensen divergence is

$$tJ_\alpha(\boldsymbol{a}, \boldsymbol{a}') \doteq \frac{1}{\sqrt{1+U^2}} \cdot J_\alpha(\boldsymbol{a}, \boldsymbol{a}') , \tag{124}$$

where $U \doteq (\varphi(\boldsymbol{a}) - \varphi(\boldsymbol{a}'))/\|\boldsymbol{a} - \boldsymbol{a}'\|_2$. There is no closed form solution for the population minimiser of tJ_α, yet we can prove the following Theorem, which builds upon Theorem 3 in [13].

Theorem 17 *In k-variates++, replace $D_t(\boldsymbol{a})$ (eq. (1)) by $D_t(\boldsymbol{a}) \doteq \min_{\boldsymbol{a}' \in \mathcal{P}} tJ_\alpha(\boldsymbol{a}, \boldsymbol{a}')$ ans suppose for simplicity that probe functions are identity: $\wp_t = \mathrm{Id}, \forall t$. Denote ϕ_{opt} the optimal noise-free potential of the clustering problem using tJ_α as distortion measure. Then there exists a* ***constant*** *$\omega > 0$ such that for* ***any*** *choice of densities $p_{\boldsymbol{\mu}_\cdot, \boldsymbol{\theta}_\cdot}$, the expected tJ_α-potential ϕ of k-variates++ satisfies:*

$$\mathbb{E}[\phi(\mathcal{A}; \mathcal{C})] \quad \leq \quad \omega \cdot \log k \cdot (6\phi_{\mathrm{opt}} + 2\phi_{\mathrm{bias}} + 2\phi_{\mathrm{var}}) \quad , \tag{125}$$

where ϕ_{var} is defined in Theorem 2 and ϕ_{bias} is defined in eq. (4).

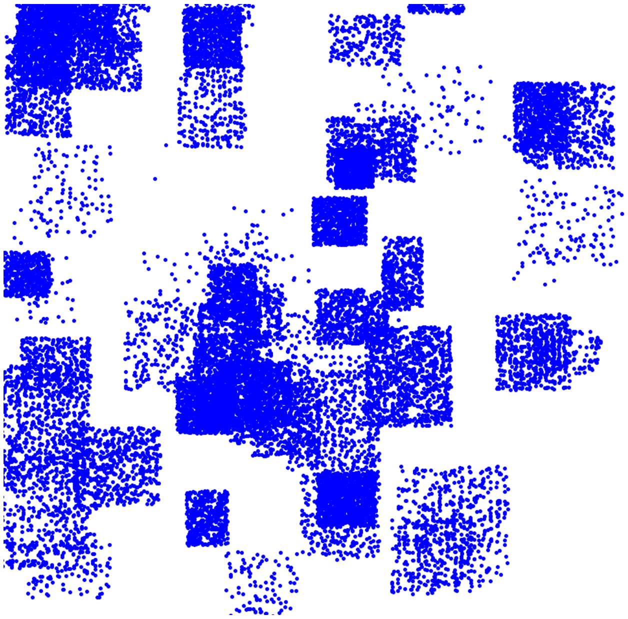
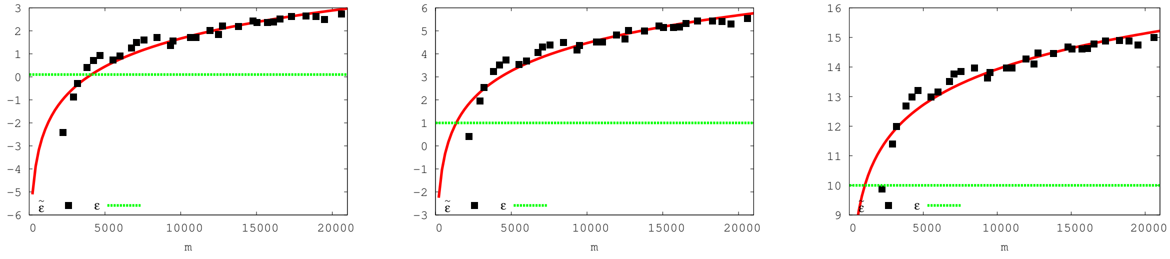


Figure 5: Final dataset for the experiments in Table 3 (plot of the two first coordinates ($d = 10$)).



$\epsilon = 0.1$ $\epsilon = 1$ $\epsilon = 10$

Table 3: Case $d = 10, k = 3$ — Plot of $\tilde{\epsilon}$ as in Theorem 12 (main file, see also eq. (126) below) and best fit for model $\tilde{\epsilon} = a + b \log m$. Figure 5 displays the final dataset obtained (see text).

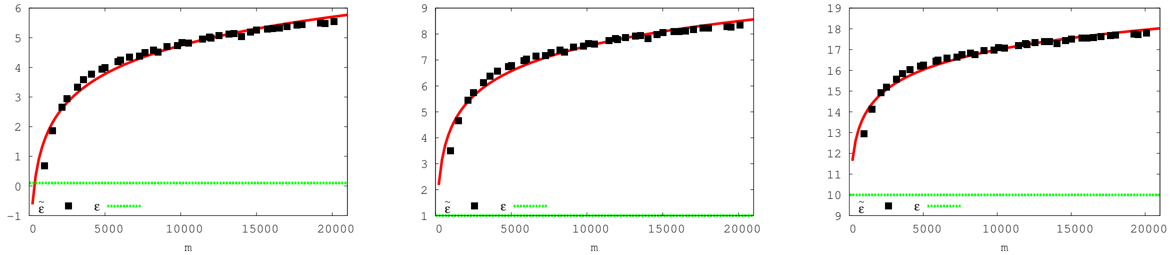
3 Supplementary Material on Experiments

3.1 Experiments on Theorem 12 and the sublinear noise regime

$\hookrightarrow$ **comments on $\tilde{\epsilon}$** An important parameter of Theorem 12 is $\tilde{\epsilon}$, which replaces ϵ in the computation of the noise standard deviation in σ_1: the larger it is compared to ϵ, the less noise we can put while still ensuring $\mathbb{P}[\mathcal{C}|\mathcal{A}']/\mathbb{P}[\mathcal{C}|\mathcal{A}] \leq \exp \epsilon$ in Definition 11 (main file). Recall its formula:

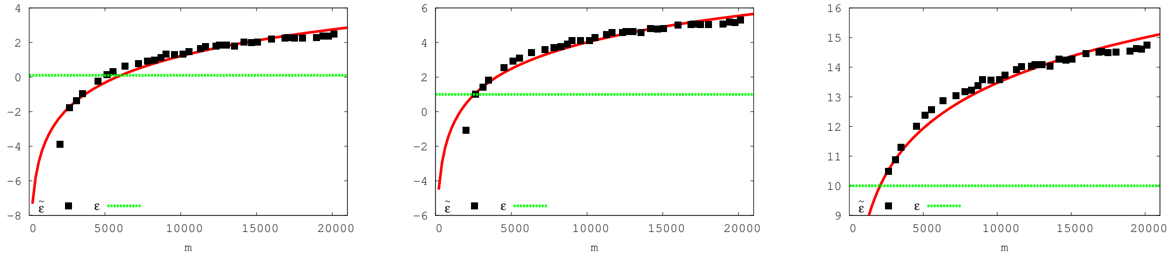
$$\tilde{\epsilon} \doteq \log \left(\frac{\exp(\epsilon) - (1 + \delta_w)^{k-1}}{f(k) \cdot \delta_w \cdot (1 + \delta_s)^{k-1}} \right) . \tag{126}$$

The experimental setting is the following one: we repeatedly sample clusters that are uniform in a subset of the domain (with limited, random size), taken to be a d-dimensional hyperrectangle of randomly chosen edge lengths. Each cluster contains a randomly picked number of points between



$\epsilon = 0.1$ $\epsilon = 1$ $\epsilon = 10$

Table 4: Case $d = 50, k = 3$ — Plot of $\tilde{\epsilon}$ as in Theorem 12 (main file, see also eq. (126) below) and best fit for model $\tilde{\epsilon} = a + b \log m$. All other parameters are the same as for Table 3.



$\epsilon = 0.1$ $\epsilon = 1$ $\epsilon = 10$

Table 5: Case $d = 50, k = 4$ — Plot of $\tilde{\epsilon}$ as in Theorem 12 (main file, see also eq. (126) below) and best fit for model $\tilde{\epsilon} = a + b \log m$. All other parameters are the same as for Table 3.

1 and 1000. After each cluster is picked, we updated an *estimation* of δ_w and δ_s:

- we compute δ_w by randomly picking $\mathcal{B}$ and $\mathcal{N}$ for a total number of n_{est} iterations, with $n_{\mathrm{est}} = 5000$;
- we compute δ_s by randomly picking $\mathcal{N}$ for a total number of n_{est} iterations. Instead of computing A then $\boldsymbol{x}$, we opt for the fast proxy which consists in replacing $\boldsymbol{c}(A)$ by a random data point, thus *without* making the $\mathcal{N}$-packed test. This should reasonably overestimate δ_s and thus slightly loosen our approximation bounds.

Figure 5 shows the dataset obtained for $d = 10$ at the end of the process. Predictably, the distribution on the whole space looks like a highly non-uniform cover by locally uniform clusters. Tables 3, 4 and 5 display results obtained for three different values of ϵ and three different values for the couple (d, k). To test the large sample regime intuition and the fact that the the noise dependence grows sublinearly in m, we have regressed in each plot $\tilde{\epsilon}$ as a function of m for

$$\tilde{\epsilon}(m) \quad = \quad a + b \log m \ . \qquad (127)$$

The plots obtained confirm a good approximation of this intuition, but they also display some more good news. The smaller ϵ, the larger can be $\tilde{\epsilon}$ relatively to ϵ, by order of magnitudes if ϵ is small.

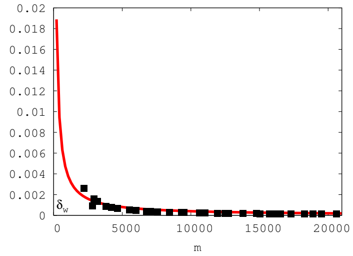
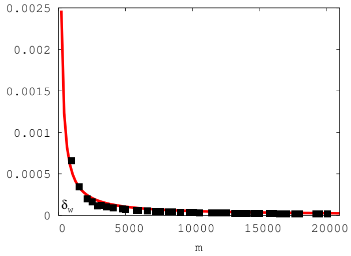
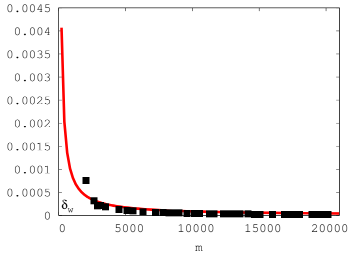
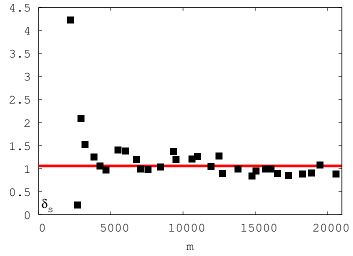
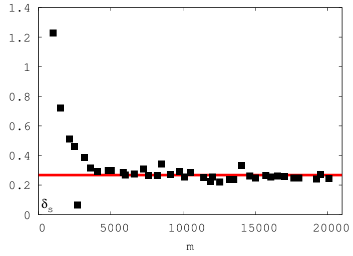
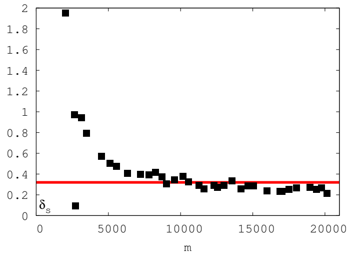
	$(d, k) = (10, 3)$	$(d, k) = (50, 3)$	$(d, k) = (50, 4)$
δ_w			
δ_s			

Table 6: Estimations of δ_w (top row) and δ_s (bottom row) as a function of m, for three values of (d, k). We also indicate the best fit for $\delta_w(m) = a/m$ (top row) and $\delta_s(m) = b$ (for $m \geq 4000$, bottom row).

Hence, despite the fact that we evetually overestimate δ_s, we still get large $\tilde{\epsilon}$. Furthermore, if k is small, this "large sample" regime in which $\tilde{\epsilon} > \epsilon$ actually happens for quite small values of m.

Also, one may remark that the curves all look like an approximate translation of the same curve. This is not surprising, since we can reformulate

$$\tilde{\epsilon} \quad = \quad \epsilon + \log\left(1 - \frac{U}{\epsilon}\right) + g(m) \ , \tag{128}$$

whene $U \doteq (1 + \delta_w)^{k-1}$ and g do not depend on ϵ. It happens that δ_w quickly decreases to very small values (bringing also a separate empirical validation of its behavior as computed in ineq. (117) in the proof of Theorem 10). Hence, we rapidly get for small m some $\tilde{\epsilon}$ that looks like

$$\begin{aligned} \tilde{\epsilon} \quad &\approx \quad \epsilon + \log\left(1 - \frac{1 + o(1)}{\epsilon}\right) + g(m) \\ &\approx \quad h(\epsilon) + g(m) \ , \end{aligned} \tag{129}$$

which may explain what is observed experimentally.

We can sumarise the global picture for $\tilde{\epsilon}$ vs ϵ by saying that it becomes more and more in favor of $\tilde{\epsilon}$ as data size (d or m) increase, but become less in favor of $\tilde{\epsilon}$ as the number of clusters k increases (predictably).

$\hookrightarrow$ **comments on δ_w and δ_s** Table 6 presents the estimated values of δ_w and δ_s for the settings of Tables 3, 4 and 5. We wanted to test the intuition as to whether, for m sufficiently large, it would hold that $\delta_w = O(1/m)$ while $\delta_s = O(1)$. The essential part is on δ_w, since such a behaviour would

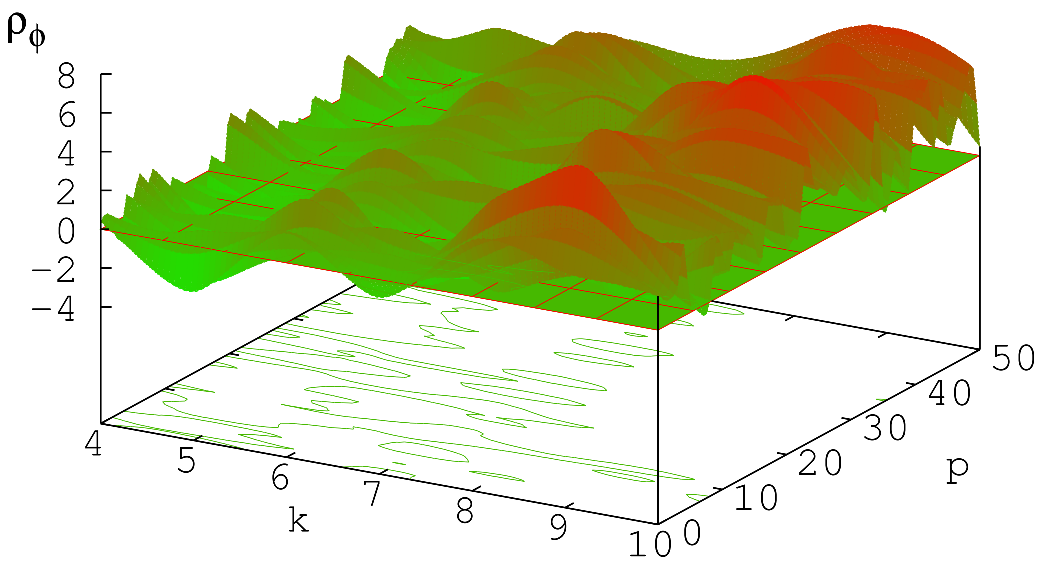


Figure 6: Simulated data — Plot of $\rho_\phi(k\text{-means++})$ as a function of k and p. Points *below* surface $z = 0$ correspond to superior performances of k-Dk-means++.

be sufficient for the sublinear growth of the noise dependence. One can check that such behaviours are indeed observed, and more: δ_w converges very rapidly to zero, at least for all settings in which we have tested data generation. Another quite good news, is that δ_s seems indeed to be $\theta(1)$, but for an actual value which is also not large, so the denominator of eq. (126) is actually driven by $f(k)$, even when, as we already said, we may have a tendency to overestimate δ_s with our randomized procedure.

3.2 Experiments with Dk-means++, k-means++ and k-means$_\parallel$

$\star$ **Experiments on synthetic data** We have generated a set of $m \approx$20 000 points using the same kind of clusters as in the experiments related to Theorem 12: we add "true" clusters until the total number of points exceeds 20 000. To simulate the spread of data among peers (Forgy nodes) and evaluate the influence of the spread of Forgy nodes (ϕ^F_s) for Dk-means++, we have devised the following protocol: let us name "true" clusters the hyperrectangle clusters used to build the dataset. Each true cluster corresponds to the data held by a peer. Then, for some $p \in [0, 100]$ (%), *each* point in *each* true cluster moves into another cluster, with probability p. The choice of the target cluster is made uniformly at random. Thus, as p increases, we get a clustering problem in which the data held by peers is more and more spread, and for which the spread of Forgy nodes ϕ^F_s

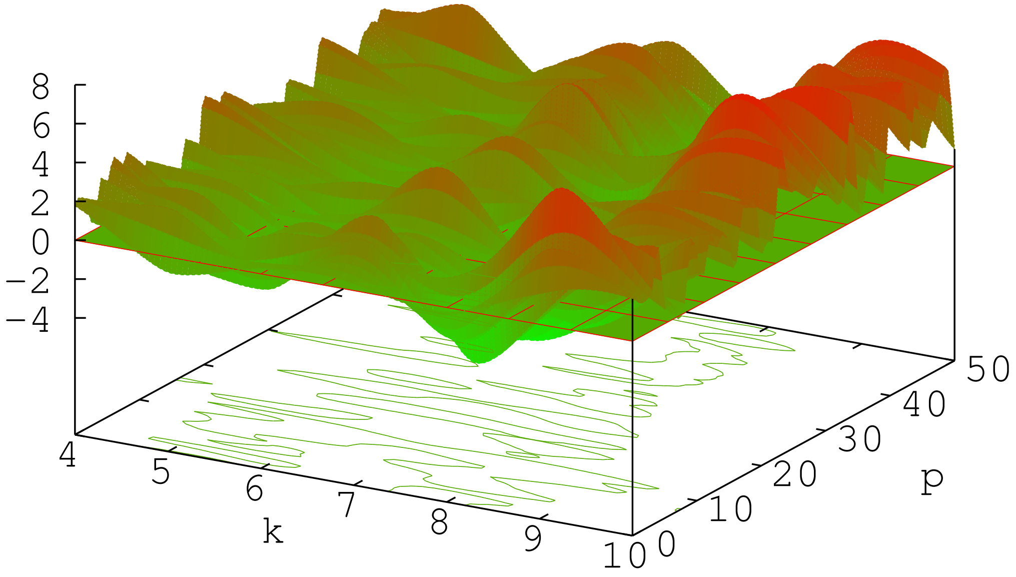


Figure 7: Simulated data — Plot of $\rho_\phi(k\text{-means}_\|)$ as a function of k and p. Points *below* surface $z = 0$ correspond to superior performances of k-Dk-means++.

increases. Figure 8 presents a typical example of the spread for $p = 50\%$. Notice that in this case many Forgy nodes have data spreading through a much larger domain than the initial, true clusters. Figure 9 displays that this happens indeed, as ϕ^F_s is multiplied by a factor exceeding 20 (compared to ϕ^F_s at $p = 0$) for the largest values of p.

We have compared Dk-means++ to k-means++ and k-means$_\|$ [4]. In the case of that latter algorithm, we follow the paper's statements and pick the number of outer iterations to be $\lceil \log \phi_1 \rceil$, where ϕ_1 is the potential for one Forgy-chosen center. We also pick $\ell = 2k$, considering that it is a value which gives some of the best experimental results in [4]. Finally, we recluster the points at the end of the algorithm using k-means++. For each algorithm $\mathcal{H} \in \{\text{k-means++}, k\text{-means}_\|\}$, we run it on the complete dataset and its results are averaged over 10 runs. We run Dk-means++ for each $p \in \{0\%, 1\%, ..., 50\%\}$. More precisely, for each p, we average the results of Dk-means++ over 10 runs. We use as metric the relative increase in the potential of Dk-means++ compared to $\mathcal{H}$:

$$\rho_\phi(\mathcal{H}) \doteq \frac{\phi(\mathrm{D}k\text{-means++}) - \phi(\mathcal{H})}{\phi(\mathcal{H})} \cdot 100 \ . \tag{130}$$

that we plot as a function of ϕ^F_s, or surface plot as a function of (k, p). The intuition for the former plot is that the larger ϕ^F_s, the larger should be this ratio, since the data held by peers spreads across the domain and each peer is constrained to pick its centers with uniform seeding.

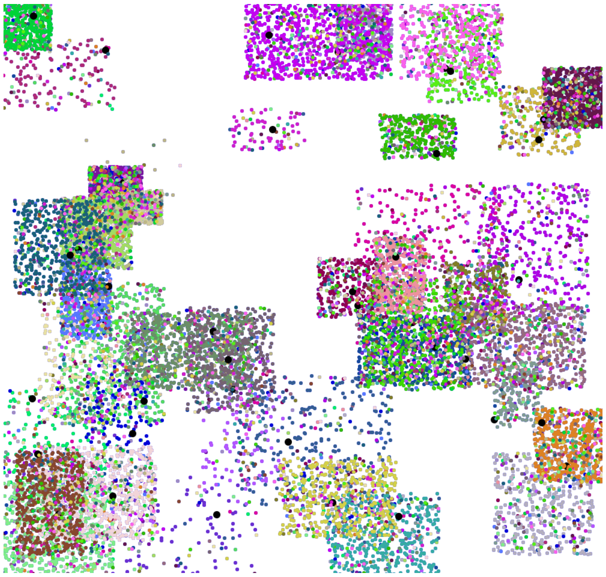
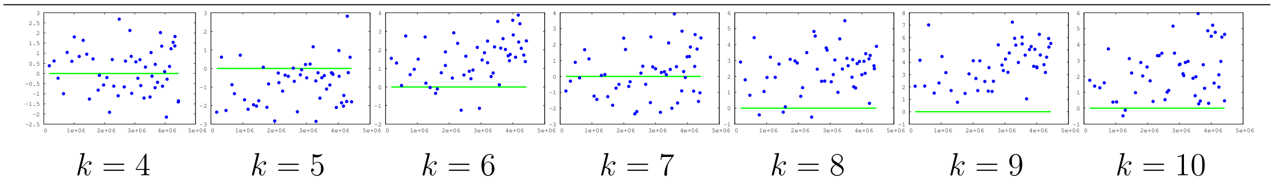


Figure 8: Example dataset obtained for $p = 50\%$ $(d = 50)$. Each color represents the points held by a peer (Forgy node) after the process of moving each point from a true cluster to another cluster with probability $p = 0.5$. Big black dots are the datapoints that are the closet to the true cluster centers.

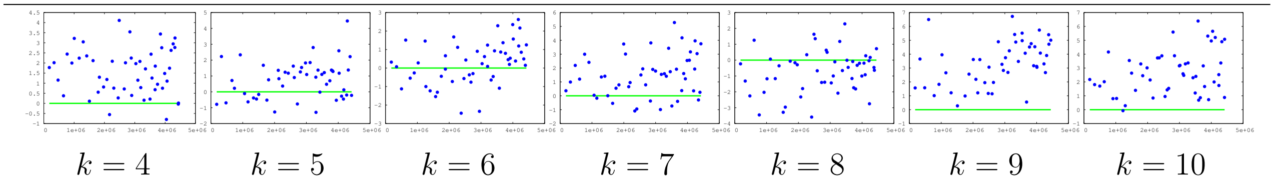
$\hookrightarrow$ **Dk-means++ vs k-means++** Figure 7 presents results for $\rho_\phi(k\text{-means++}) = f(\phi_s^F)$ obtained for various k. First, the intuition is indeed confirmed for $k = 8, 9, 10$, but an interesting phenomenon appears for $k = 5$: Dk-means++ almost consistently beats k-means++. The decrease in the average potential ranges up to 3%. Furthermore, this happens even for large values of ϕ_s^F. Finally, for *all but one* value of k, there exists spread values for which Dk-means++ beats k-means++. The surface plot in Figure 6 displays that superior performances of Dk-means++ are probably not random. One possible explanation to this phenomenon relies on the expression of ϕ_{bias} given in the proof of Theorem 4 (eq. (21)), recalled here:

$$\begin{aligned}\phi_{\text{bias}} &\doteq \sum_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{\mu}_{\boldsymbol{a}} - \boldsymbol{c}_{\text{opt}}(\boldsymbol{a})\|_2^2 \\ &= \sum_{i\in[n]} \sum_{\boldsymbol{a}\in\mathcal{A}_i} \|\boldsymbol{c}(\mathcal{A}_i) - \boldsymbol{c}_{\text{opt}}(\boldsymbol{a})\|_2^2 \ .\end{aligned} \tag{131}$$



$k = 4$ | $k = 5$ | $k = 6$ | $k = 7$ | $k = 8$ | $k = 9$ | $k = 10$

Table 7: Simulated data — Plot of ratio $\rho_\phi(k\text{-means++})$ in eq. (130) as a function of ϕ_s^F. Points *below* the green line correspond to (average) runs in which Dk-means++ *beats* k-means++.



$k = 4$ | $k = 5$ | $k = 6$ | $k = 7$ | $k = 8$ | $k = 9$ | $k = 10$

Table 8: Simulated data — Plot of ratio $\rho_\phi(k\text{-means}_{\|})$ in eq. (130) as a function of ϕ_s^F. Points *below* the green line correspond to (average) runs in which Dk-means++ *beats* k-means$_{\|}$.

Recall that ϕ_{bias} can be $< \phi_{\text{opt}}$, and it can even be zero, in which case Theorem 2 says that the approximation bound may actually be *better* than that of k-means++ in [1] (furthermore, $\eta = 0$ for Dk-means++). Hence, what happens is pobably that in several cases, there exists a union of peers data (the number of peers is larger than k) that gives a at least reasonably good approximation of the global optimum. In all our experiments indeed, we obtained a number of peers larger than 30.

$\hookrightarrow$ **Dk-means++ vs k-means$_{\|}$** Figure 7 appear to display performances for Dk-means++ that are even more in favor of Dk-means++, compared to k-variates++. Figure 8 presents results for $\rho_\phi(k\text{-means}_{\|}) = f(\phi_s^F)$ obtained for various k. The fact that each of them is a vertical translation of a picture in Figure 7 comes from the fact that the results of k-means$_{\|}$ and k-means++ do not depend on the spread of the neighbors ϕ_s^F.

$\star$ **Experiments on real world data** We consider the EuropeDiff dataset[2] (Dataset characteristics provided in Table 9). Figures 10 and 11 give the results for the equivalent settings of the experimental data. To simulate N peers with real data, reasonably spread geographically, we have sampled N points (”peer centers”) with k-means++ seeding in data and then aggregated for each peer the subset of data in the corresponding Voronoi 1-NN cell. We then simulate the spread for parameter p as in the simulated data. Figures 10 and 11 globally display (and confirm) the same trends as for the simulated data. They, however, clearly emphasize this time that the spread of Forgy nodes ϕ_s^F is one key parameter that drives the performances of Dk-means++. Notice also that Dk-means++ remains on this dataset competitive up to $p \geq 30\%$, which means that it remains competitive when a significant proportion of peers' data is scattered without any constraint.

To further address the way the spread of Forgy nodes affects results, we have used another real

[2] http://cs.joensuu.fi/sipu/datasets/

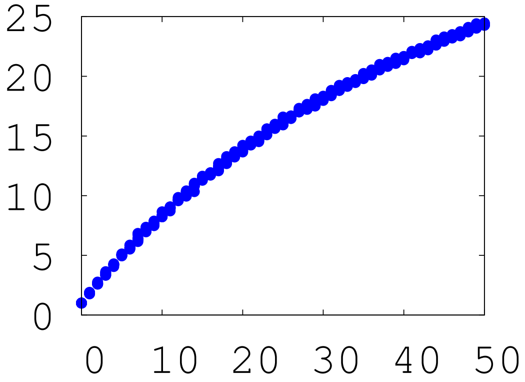


Figure 9: Simulated data — Relative increase of spread, $\phi_s^F(p)/\phi_s^F(0)$, through the runs, as a function of p.

world data with highly non-uniform distribution, Mopsi-Finland locations[3] ($m = 13467, d = 2$). We have sampled peers using two different schemes for the peer centers: k-means++ and Forgy. In this latter initialisation, we just pick peer centers at random. In the former k-means++ initialisation, the initial peer centers are much more evenly geographically spread before we complete the peers data with the closest points. They remain more spread after the $p\%$ uniform displacement of data between peers, as shown on the top plots of Figure 12. What is interesting about this data is that it displays that if peers' data are indeed geographically located, then Dk-means++ is competitive up to quite reasonable values of $p \leq 20\%$ (depending on k). That, is Dk-means++ works well when each peer aggregates 80 $\%$ data which is reasonably "localized in the domain" and 20 $\%$ data which can be located everywhere in the domain.

3.3 Experiments with k-variates++ and GUPT

Among the state-of-the-art approaches against which we could compare k-variates++, there are two major contenders, PINQ [14] and GUPT [15]. Even when PINQ is a broad system, we switched our preferences to GUPT for the following reasons. The performance of k-means based on PINQ relies on two principal factors: the initialisation (like in the non differentially private version) and the number of iterations. To compete against heavily tuned specific applications, like k-variates++, this scheme requires substantial work for its optimisation. For example, if one allocates part of the privacy budget to release a differential private initialisation, the noise has to be proportional to the domain width, which would release poor centers. Also, generating points uniformly at random from the domain, to obtain data-independent initial centers, yields to a poor initialisation. Finally, the number of iterations has to be tuned very carefully: if too small, the algorithm keeps poor solutions; if too large, the number of iteration increase the added noise for privacy and harms PINQ's final accuracy. We thus chose GUPT. k-means implemented in the GUPT proceeds the following way: the dataset is cut in a certain number of blocks ℓ (following [15], we fix $\ell = m^{0.4}$ in our experiments), the usual k-means algorithm is performed on each block. Before releasing the final centroids, results are aggregated and a noise is applied. Finally, we also compare against

[3] http://cs.joensuu.fi/sipu/datasets/

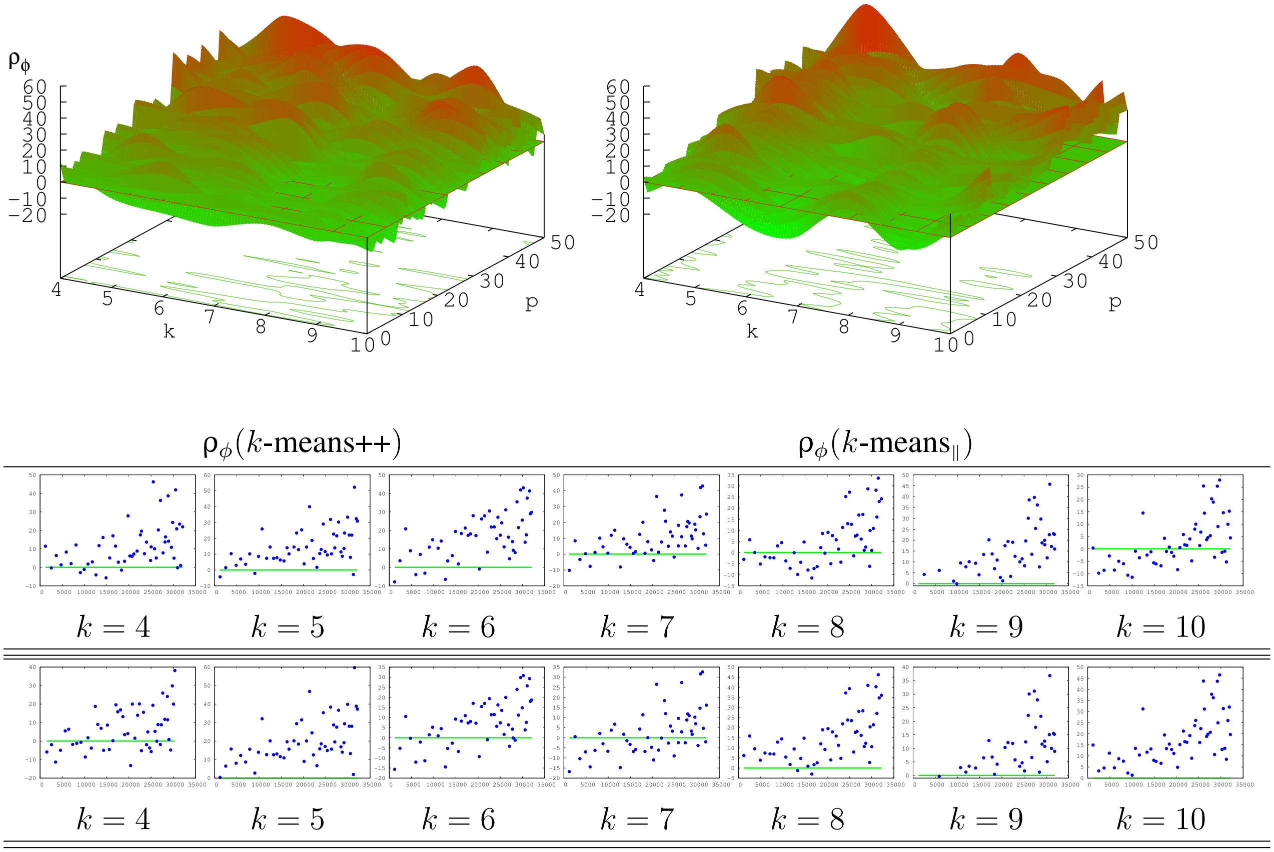


Figure 10: Experiments on real world data ”EuropeDiff” with $N = 30$ simulated peers. Top plot: Plots corresponding to Figures 6 (left) and 7 (right). Middle and bottom plot ranges: plots corresponding respectively to Figures 7 and 8.

the vanilla approach of *Forgy Initialisation* using the Laplace mechanism. The noise rate (*i.e.*, standard deviation) is then proportional to $\propto kR/\epsilon$ (we do not run k-means afterwards, hence the privacy budget remains “small”). In comparison, GUPT adds noise $\propto kR/(\ell\epsilon)$ at the end of this aggregation process. Note that we disregard the fact that our data are multidimensional, which should require a finer-grained tuning of ℓ, and choose to rely on the $\ell = m^{0.4}$ suggestion from [15].

$\hookrightarrow$ **Comparison on real world domains** Our domains consist of 3 real-world datasets[4]. Lifesci contains the value of the top 10 principal components for a chemistry or biology experiment. Image is a 3D dataset with RGB vectors, and finally EuropeDiff is the differential coordinates of Europe map.

Table 9 presents the extensinve results obtained, that are averaged in the main file. We have fixed $\epsilon = 1$ in the differentially privacy parameters. The column $\tilde{\epsilon}$ (eq. (18) in the main file) provides the differential privacy parameter which is equivalent from the protection standpoint, but exploits the computation of δ_w, δ_s (which we compute exactly, and not in a randomized way like in

[4] http://cs.joensuu.fi/sipu/datasets/

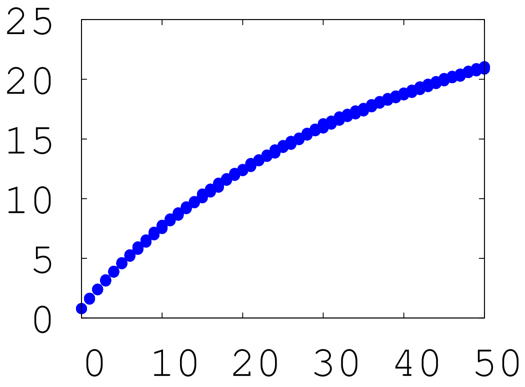


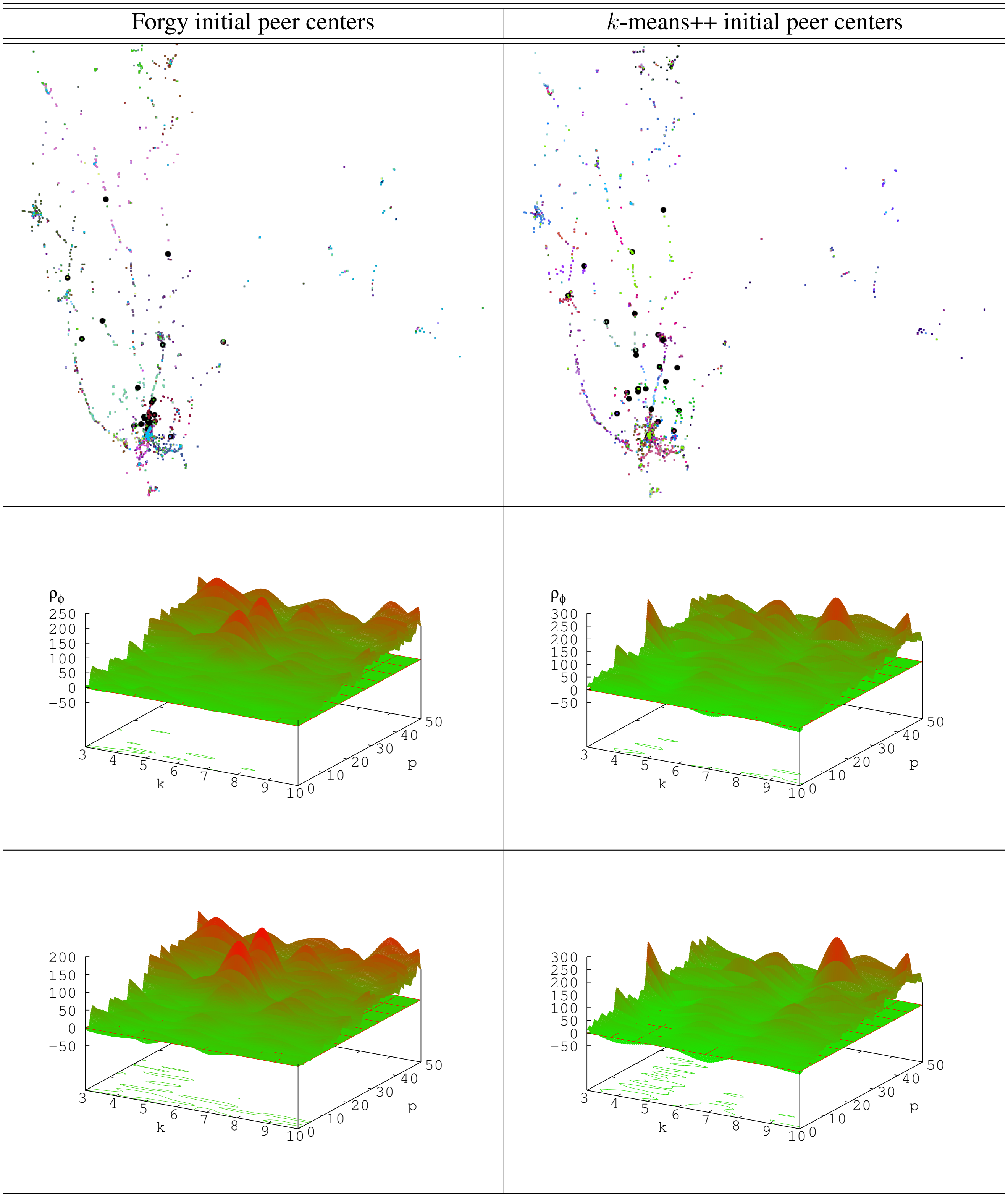
Figure 11: Experiments on real world data ”EuropeDiff” with $N = 40$ simulated peers. Plot corresponding to Figure 9.

the experiments on Theorem 12 above) and ineq. (58). Therefore, each time $\tilde{\epsilon} > \epsilon$ (=1 in our applications), it means that our analysis brings a sizeable advantage over “raw protection” by Laplace mechanism (in our application we chose for $p_{\boldsymbol{\mu}_a,\boldsymbol{\theta}_a}$ a Laplace distribution). R is computed from the data by an upperbound of the smallest enclosing ball radius. The results display several interesting patterns. First, the largest the domain, the better we compare with respect to the other algorithms. On EuropeDiff for example, we often have the ratio of the potentials $\phi(\mathrm{GUPT})/\phi(k\text{-variates++})$ of the order of *dozens*. Also, the performances of k-variates++ degrade if k increases, which is again consistent with the “good” regime of Theorem 10.

$\hookrightarrow$ **Comparison on synthetic domains** The synthetic datasets contain points uniformly sampled on a unit d-ball, in low dimension $d = 2$ and higher dimension $d = 15$, we generated datasets with size in $\{10^5, 10^6\}$.

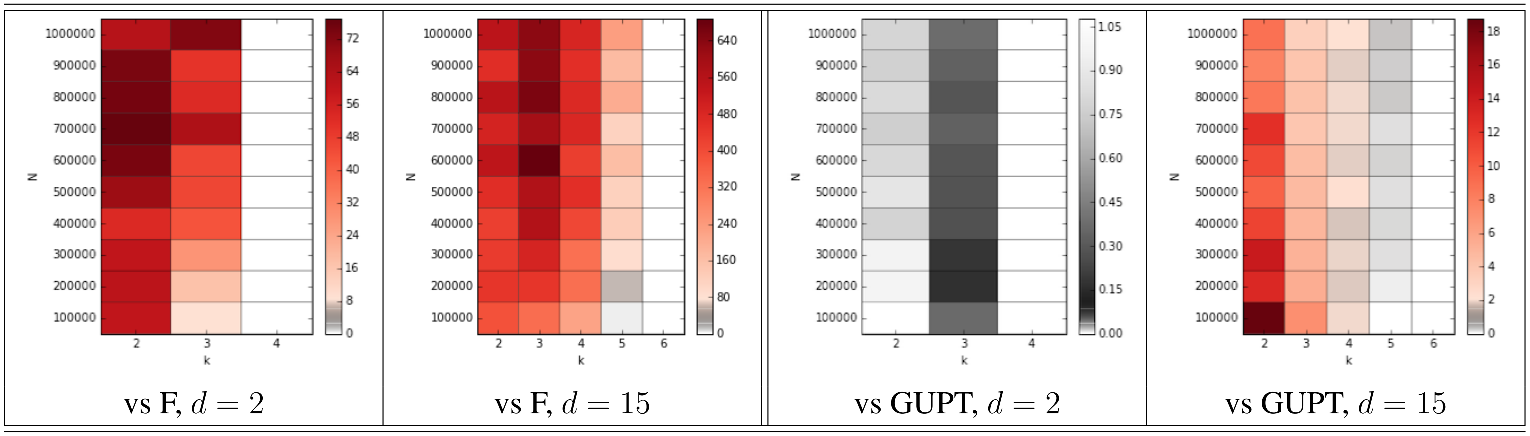
Dataset	m	d	k	$\tilde{\epsilon}$	ρ'_ϕ(F-DP)	ρ'_ϕ(GUPT)
LifeSci	26733	10	2	8.5	311	1.6
			3	4.4	172	0.4
			4	0.6	6	0.02
Image	34112	3	2	12.6	300	4.8
			3	3.2	77	0.9
EuropeDiff	169308	2	2	19.0	1200	46.1
			3	21.0	3120	66.5
			4	18.0	3750	55.0
			5	14.0	4000	51.0
			6	10.4	5000	36.0
			7	6.6	2600	26.0
			8	1.8	350	2.0

Table 9: Comparison of k-variates++, Forgy Initialisation differentially private (F-DP) and GUPT on the real world domains (results averaged in the main file)). On each domain, we compute ratio ρ'_ϕ of the clustering potential of the contender to that of k-variates++, a value > 1 indicating that k-variates++ is better. The potential of each algorithm has been averaged over 30 runs. $\tilde{\epsilon}$ is given in eq. (18) (main file).


Forgy initial peer centers

k-means++ initial peer centers

Figure 12: Mopsi-Finland locations data — Top: peer centers (big black dots) after $p = 50\%$ moving probability changes in data. Remark from the right plot (k-means++ initial peer centers) that peer data are less ”attracted” towards the highest density regions. Center: plots of $\rho_{\phi}(k\text{-means++})$. Bottom: plots of $\rho_{\phi}(k\text{-means}_{\|})$.



vs F, $d = 2$

vs F, $d = 15$

vs GUPT, $d = 2$

vs GUPT, $d = 15$

Figure 13: k-variates++ vs Forgy initialisation differentially private and GUPT. We use ratio ρ'_ϕ between the potential of the contender in (F-DP, GUPT) over the potential of k-variates++ (potentials are averaged 30 times). The more red, the better is k-variates++ with respect to the contender. Grey values indicate less positive outcomes for k-variates++; white values indicate that k-variates++ does not manage to find an ϵ' larger than ϵ, and thus does not manage to put smaller noise rate than in the Laplace mechanism.

References

[1] D. Arthur and S. Vassilvitskii. k-means++ : the advantages of careful seeding. In *19th SODA*, pages 1027 – 1035, 2007.

[2] Y. Wang, Y.-X. Wang, and A. Singh. Differentially private subspace clustering. In *NIPS*28*, 2015.

[3] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *40th ACM STOC*, pages 75–84, 2007.

[4] B. Bahmani, B. Moseley, A. Vattani, R. Kumar, and S. Vassilvitskii. Scalable k-means++. In *38th VLDB*, pages 622–633, 2012.

[5] M.-F. Balcan, S. Ehrlich, and Y. Liang. Distributed k-means and k-median clustering on general communication topologies. In *NIPS*26*, pages 1995–2003, 2013.

[6] N. Ailon, R. Jaiswal, and C. Monteleoni. Streaming k-means approximation. In *NIPS*22*, pages 10–18, 2009.

[7] E. Liberty, R. Sriharsha, and M. Sviridenko. An algorithm for online k-means clustering. *CoRR*, abs/1412.5721, 2014.

[8] C. Dwork and A. Roth. The algorithmic foudations of differential privacy. *Found. & Trends in TCS*, 9:211–407, 2014.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *3rd TCC*, pages 265–284, 2006.

[10] S. Jegelka, S. Sra, and A. Banerjee. Approximation algorithms for tensor clustering. In *20th ALT*, pages 368–383, 2009.

[11] R. Nock, P. Luosto, and J. Kivinen. Mixed Bregman clustering with approximation guarantees. In *19th ECML*, pages 154–169, 2008.

[12] R. Nock, F. Nielsen, and S.-I. Amari. On conformal divergences and their population minimizers. *IEEE Trans. IT*, 62:1–12, 2016.

[13] F. Nielsen and R. Nock. Total Jensen divergences: definition, properties and clustering. In *40th IEEE ICASSP*, pages 2016–2020, 2015.

[14] F. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. *Communications of the ACM*, 53(9):89–97, 2010.

[15] P. Mohan, A. Thakurta, E. Shi, D. Song, and D.-E. Culler. GUPT: privacy preserving data analysis made easy. In *38th ACM SIGMOD*, pages 349–360, 2012.