Coresets for Nonparametric Estimation — the Case of DP-Means

Olivier Bachem OLIVIER.BACHEM@INF.ETHZ.CH
Mario Lucic LUCIC@INF.ETHZ.CH
Andreas Krause KRAUSEA@ETHZ.CH
ETH Zurich, Switzerland

Abstract

Scalable training of Bayesian nonparametric models is a notoriously difficult challenge. We explore the use of *coresets* – a data summarization technique originating from computational geometry – for this task. Coresets are weighted subsets of the data such that models trained on these coresets are provably competitive with models trained on the full dataset. Coresets sublinear in the dataset size allow for fast approximate inference with provable guarantees. Existing constructions, however, are limited to *parametric* problems. Using novel techniques in coreset construction we show the existence of coresets for DP-Means – a prototypical nonparametric clustering problem – and provide a practical construction algorithm. We empirically demonstrate that our algorithm allows us to efficiently trade off computation time and approximation error and thus scale DP-Means to large datasets. For instance, with coresets we can obtain a computational speedup of $45\times$ at an approximation error of only 2.4% compared to solving on the full data set. In contrast, for the same subsample size, the "naive" approach of uniformly subsampling the data incurs an approximation error of 22.5%.

1. Introduction

Traditional models in machine learning often require an explicit choice of the model capacity via hyperparameters. For example, in K-Means clustering, the number of clusters must be selected *a priori* even though this quantity is not known for many practical applications. The standard remedy to this problem is to consider models of varying capacities and to perform model selection using criteria such as AIC or BIC.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

Instead of comparing models of different complexity, Bayesian nonparametric models infer the complexity from the data. The fact that the model complexity can scale with the amount of observed data enables one to fit a single model to datasets of varying sizes, making nonparametric models especially appealing in the age of Big Data. This flexibility is a consequence of Bayesian nonparametric models having an infinite-dimensional parameter space (Orbanz & Teh, 2010). At the same time a finite number of observations can be fully explained by only a finite subset of parameters ensuring tractability of the problem. Of course, this flexibility comes at a cost: Training nonparametric models on massive data sets is notoriously difficult.

As a concrete example, consider the *Dirichlet process (DP) mixture model* (Antoniak, 1974) which is a widely used Bayesian nonparametric model for clustering. It assumes that the i-th observation depends on a parameter θ_i which is sampled from a Dirichlet process. Since Dirichlet processes are almost surely discrete, the model can be viewed as a mixture model with an unbounded number of cluster centers. If the base distribution of the Dirichlet process and the distribution of an observation i given θ_i are both Gaussian, we obtain the Gaussian Dirichlet process mixture which can be viewed as an unbounded extension of the Gaussian mixture model. Recently, Kulis & Jordan (2012) apply the technique of *small variance asymptotics* to the Gibbs sampler of a Gaussian Dirichlet process mixture model. This leads to a hard clustering algorithm that monotonically decreases an object function similar to the K-Means objective. In contrast to K-Means, DP-Means clustering allows solutions with an arbitrary number of clusters, but introduces a penalty term proportional to the number of clusters used in the solution. One motivation for DP-Means instead of DP mixtures is that inference becomes more efficient. Nevertheless, scaling it to massive datasets remains a difficult challenge.

Related Work. In case of DP mixtures, the traditional approaches are based on MCMC sampling (MacEachern, 1994; Escobar & West, 1995; Neal, 2000). Yet, they can

be slow to converge which prevents them from being applied in large-scale settings. As a result there has been an increased interest in alternative inference methods for Bayesian nonparametrics. In the case of Dirichlet process mixtures the most popular approaches include recursive approximations (Newton & Zhang, 1999), sequential importance sampling (MacEachern et al., 1999) and variational Bayes (Blei & Jordan, 2006).

Our Contributions. We investigate a novel approach to scaling Bayesian nonparametrics. We apply coresets – a technique from computational geometry – that allows one to efficiently summarize a large dataset by a small weighted subset of representative points. Coresets come with strong theoretical guarantees on how well they approximate the original point set while usually being sublinear, if not independent, in the number of observations. This allows for fast approximate inference for large datasets by running potentially slow algorithms on small coresets. However, to the authors' knowledge, all previous work on coresets was focused exclusively on parametric models such as K-Means (Har-Peled & Mazumdar, 2004; Har-Peled & Kushal, 2005; Feldman et al., 2007; Chen, 2009; Langberg & Schulman, 2010; Feldman & Langberg, 2011a) and Gaussian mixtures (Feldman et al., 2011). We extend the existing coreset theory and show that it is possible to apply coresets to nonparametric models. DP-Means clustering is an ideal target for our approach as it is similar to K-Means clustering, yet nonparametric in nature. Due to the unbounded parameter space, coresets are not trivially applied to nonparametric models. Several steps of the standard coreset construction methodology require adaptation and development of new techniques. The existence and successful construction of coresets for DP-Means is thus a first step to building coresets for a more general class of Bayesian nonparametric models. Similarly, coresets have first been developed for hard clustering problems such as K-Means and then later extended to more advanced models such as Gaussian mixtures.

As our key contributions, we:

- theoretically prove the existence of coresets even for unbounded queries,
- propose a practical coreset construction algorithm that retains strong theoretical guarantees, and
- empirically validate our theoretical findings and demonstrate the effectiveness of our algorithm compared to full inference and uniform subsampling.

2. Background

DP-Means clustering It is known that LLoyd's algorithm may be derived as the limit of the expectation-maximization algorithm used to fit Gaussian mixture models as the variance approaches zero (Kulis & Jordan, 2012). Similarly, Kulis & Jordan (2012) consider the limit of the Gaussian Dirichlet process mixture model where both cluster centers and the individual data points are sampled from a multivariate Gaussian. More specifically, each data point x_i is sampled from a spherical Gaussian with mean μ_i and variance σ where the means are again sampled from a Dirichlet process with a concentration parameter α and a spherical Gaussian prior with variance ρ as its base distribution. An equivalent but more intuitive interpretation of this generative model assumes that a potentially infinite number of cluster centers are generated from a Gaussian prior with variance ρ. Cluster assignments of n data points are then sampled using a *Chinese Restaurant Process* with concentration parameter α. Each data point is finally generated from a Gaussian with the assigned cluster center as its mean and variance σ.

By analyzing the case where σ approaches zero, Kulis & Jordan (2012) show that the Gibbs sampler of Neal (2000) converges to a hard clustering algorithm. Essentially, this algorithm strongly resembles LLoyd's algorithm for K-Means, except that the number of cluster centers is not fixed. Instead, during the assignment step, a point is assigned to the closest center only if it is within distance λ from the closest center, otherwise it is chosen as a new cluster center. The algorithm is deterministic as the initial solution consists of one cluster center at the global mean of the data and it monotonically decreases the following objective function (also called the *DP-Means cost function*)

$$\mathrm{cost}_{DP}(\mathcal{P}, Q) = \sum_{p \in \mathcal{P}} w_p \min_{q \in Q} \mathrm{dist}(p, q)^2 + |Q|\lambda$$

where $\lambda > 0$, $\mathcal{P}$ is a weighted set of n points in $\mathbb{R}^d$ and $Q \subset \mathbb{R}^d$ a non-empty set of cluster centers.

This cost function gives rise to the *DP-Means clustering problem* where the goal is to find a finite, non-empty set of cluster centers $Q \subset \mathbb{R}^d$ minimizing the DP-Means objective function. The first term of the objective is identical to the quantization error in K-Means and measures how well the clusters approximate the data. The second term is a penalization term that is proportional to the number of cluster centers used in Q.

The DP-Means clustering problem can be viewed as a nonparametric extension of the K-Means objective since an arbitrary number of cluster centers may be used. The model complexity is inferred from the data via a tradeoff between the number of cluster centers used and the quality of the clustering. Interestingly, the DP-Means objective has other natural motivations such as the Elbow method and, as noted in Kulis & Jordan (2012), is related to objective functions studied in connection with the Akaike Information Criterion.

As with K-Means clustering, it is challenging to solve the DP-Means clustering problem for datasets where the number of samples is prohibitively large. A "naive" approach is to solve the clustering problem on a random subset of the data with the hope that the solution on this subset is close to the solution on the full dataset. However, uniform subsampling provides no theoretical guarantees and one can easily construct examples where this approach is guaranteed to fail. For example, consider K-Means clustering with $k = 2$ applied to two well separated clusters where the first cluster consists of $\log n$ data points. A finite uniform subsample of the data will, with high probability, include only points from the second cluster and the solution on the subsample will produce two centers in the second cluster. Hence, by increasing the distance between the two clusters, the objective value can be made arbitrarily large. Coresets introduced in the next section provide a remedy for this issue.

Coresets A coreset is a weighted subset of the data such that the quality of any clustering evaluated on the coreset closely approximates the true quality on the full dataset. Consider a cost function depending on a set of points $\mathcal{P}$ and a query or solution $Q \in \mathcal{Q}$ that is additively decomposable into non-negative functions $\{f_Q(p)\}_{p\in\mathcal{P}}$, i.e.

$$\mathrm{cost}(\mathcal{P}, Q) = \sum_{p\in\mathcal{P}} f_Q(p).$$

The idea of coresets is to find a weighted subset $\mathcal{C}$ such that the cost of a query Q can be approximated on $\mathcal{C}$ by

$$\mathrm{cost}(\mathcal{C}, Q) = \sum_{(w,c)\in\mathcal{C}} w f_Q(c).$$

The notion of approximation is formalized as the coreset property: A weighted subset $\mathcal{C}$ is an ϵ-coreset of $\mathcal{P}$ if it approximates the cost function of the full dataset up to a multiplicative factor of $1 \pm \epsilon$ uniformly for all queries $Q \in \mathcal{Q}$, i.e.,

$$|\mathrm{cost}(\mathcal{P}, Q) - \mathrm{cost}(\mathcal{C}, Q)| \leq \epsilon\, \mathrm{cost}(\mathcal{P}, Q).$$

We note that since the cost contributions $f_Q(p)$ and the possible space of solutions $\mathcal{Q}$ depend on the considered problem, coresets are inherently problem-specific.

The main motivation for constructing coresets is to approximately solve optimization problems by running any solver on the coreset instead of the full dataset. The idea is that since the coreset property bounds the approximation error for all queries, the difference between the solution on the full dataset and the solution on the coreset is bounded. Furthermore, coresets are generally small, i.e., sublinear in, or even independent of, the number of samples. This allows one to apply optimization algorithms with higher computational complexity, such as squared or cubic dependence on the number of samples, making coresets an ideal choice for computationally hard problems. For K-Means, coresets even allow for a polynomial time approximation scheme (Feldman et al., 2007).

Additionally, coresets are a practical and flexible tool that requires no assumptions on the data. While the theory behind coresets is very technical and requires elaborate tools from computational geometry to prove the strong theoretical guarantees, the resulting coreset construction algorithms are simple to implement.

A key property of coresets is that they can be constructed both in a distributed and a streaming setting. The constructions rely on the property that both unions of coresets and coresets of coresets are coresets (Har-Peled & Mazumdar, 2004) – albeit with different ϵ. Feldman et al. (2011) use these properties to construct coresets in a tree-wise fashion which can be parallelized in a Map-Reduce style or used to maintain an up-to-date coreset in a streaming setting.

3. Coresets for DP-Means

Our main contribution is the first construction of coresets for a class of nonparametric models: DP-Means.

Definition 3.1. Let $\epsilon > 0$ and $\mathcal{P}$ be a set of n points in $\mathbb{R}^d$. The weighted set $\mathcal{C}$ is an $(\epsilon, \bar{k})$-*coreset* for the DP-Means clustering of $\mathcal{P}$ if for any query, i.e. any non-empty set Q, of at most $\bar{k}$ centers in $\mathbb{R}^d$

$$|\mathrm{cost}_{DP}(\mathcal{P}, Q) - \mathrm{cost}_{DP}(\mathcal{C}, Q)| \leq \epsilon\, \mathrm{cost}_{DP}(\mathcal{P}, Q).$$

If this property holds with $\bar{k} = \infty$, the weighted set $\mathcal{C}$ is called an ϵ-*coreset*.

This definition already highlights the challenge of applying coresets to the nonparametric setting. The size of queries Q is not fixed as in the parametric setting and the coreset property has to hold uniformly for all query sizes. For the case of $\bar{k} = \infty$, queries can even be of unbounded size.

3.1. Theoretical existence result using exponential grids

Our first result shows the existence of ϵ-coresets for the DP-Means clustering problem that are sublinear in the number of data points n if the optimal number of centers k^* is sublinear in n. Naturally, if the optimal number of centers is linear in n, then no coreset sublinear in n can exist.

Theorem 3.2. Let $0 < \epsilon \leq 1$ and let $\mathcal{P}$ be a set of n points in $\mathbb{R}^d$. Then there exists an ϵ-coreset for the DP-Means clustering of $\mathcal{P}$ with size $\mathcal{O}\Big(\frac{d^d k^* \log n}{\epsilon^d}\Big)$ where k^* is the optimal number of centers.

Proof sketch. This result is obtained by applying the exponential grid approach of Har-Peled & Mazumdar (2004) to the DP-Means cost function. Assume that the optimal solution to the DP-Means clustering problem is known.

One can then build an exponential grid around each of the cluster centers and project all data points in a grid cell to an arbitrary representative. It can be shown that the number of grid cells is $\mathcal{O}\Big(\frac{d^d k^* \log n}{\epsilon^d}\Big)$ and that the sum of the cost differences induced by the projection is bounded by $\epsilon \operatorname{cost}_{DP}(\mathcal{P}, Q)$ implying the required result. The full proof can be found in the Supplementary Materials.

Remarkably, this result proves that coresets of *sublinear size* exist for queries of unbounded size. While this is an encouraging theoretical result, this coreset construction is not practical. Implementing the construction using exponential grids is tedious and the coreset size has an exponential dependence on the ambient dimension. Moreover, this construction assumes knowledge of the optimal solution – the very thing we aim to compute. Nevertheless, in the following section we show a practical coreset construction using importance sampling.

3.2. Practical coresets using importance sampling

Our practical coreset construction scheme builds upon the framework by Feldman & Langberg (2011a). The idea is to first find a rough approximation (*bicriteria approximation*) to the optimal solution and then use this solution to calculate a non-uniform sampling distribution for the data points. The coreset is obtained by sampling a sufficient number of points from this distribution and setting the weights of the points inversely proportional to the sampling probabilities. The intuition is that any importance sampling scheme produces an unbiased estimator and that the variance of this estimator can be bounded. Using the theory of ϵ-approximators from computational geometry it can then be shown that for enough samples this leads to the required coreset property.

Constructing a rough approximation to the optimal solution of a nonparametric model is a non-trivial task. For K-Means it is usually found by relaxing the clustering problem such as allowing the use of more than k cluster centers. This relaxation is not possible for DP-Means as the number of cluster centers used has a direct impact on the objective function. It is even harder as the rough approximation needs to be nonparametric and itself infer a reasonable number of cluster centers.

For a DP-Means problem instance defined by a set of points $\mathcal{P}$ in $\mathbb{R}^d$ and a hyperparameter $\lambda > 0$, we propose the following coreset construction which is illustrated in Figure 1. It consists of three steps:

Step 1 To find a (rough) approximation of the optimal solution we propose the algorithm *DP-Means++* (Algorithm 1) which is inspired by the seeding step of K-Means++ (Arthur & Vassilvitskii, 2007). K-Means++ selects the initial cluster centers using k rounds of D^2-sampling where the first cluster center is sampled uniformly and additional points are then sampled with probability proportional to the minimum squared distance to the already selected cluster centers. DP-Means++ also uses D^2-sampling but with a critical twist - the number of centers sampled is not fixed but inferred from the data using a stopping condition. Intuitively, this stopping condition manages the tradeoff between quantization error and penalization term which is the essential challenge of DP-Means clustering. Algorithm 1 produces solutions of size k' and is $\mathcal{O}(\log k')$ competitive to the optimal solution. Furthermore, it can be shown that Algorithm 1 also provides us with an upper bound $\bar{k} = k'(16(\log_2 k' + 2) + 1)$ on the optimal number of cluster centers k^*.

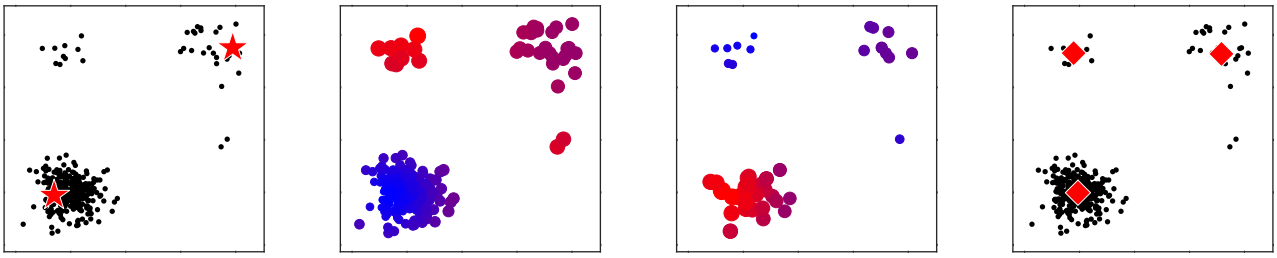
Algorithm 1 DP-Means++

Require: Set of data points $\mathcal{P}$, parameter λ
  Uniformly sample $a \in \mathcal{P}$ and set $A = \{a\}$
  while $\sum_{p \in \mathcal{P}} \operatorname{dist}(p, A)^2 > 16\lambda|A|(\log_2 |A| + 2)$ **do**
    Sample point $a \in \mathcal{P}$ with probability $m(a) = \frac{\operatorname{dist}(a,A)^2}{\sum_{p' \in \mathcal{P}} \operatorname{dist}(p',A)^2}$ and add it to A
  Return approximate solution A of cardinality k'

Algorithm 2 Importance sampling scheme

Require: Set of data points $\mathcal{P}$, approximate DP-Means solution A of cardinality k'
  $\alpha \leftarrow 16(\log_2 k' + 2) + 2$
  $\bar{c} \leftarrow \operatorname{cost}_{DP}(\mathcal{P}, A)/|\mathcal{P}|$
  for $a \in A$ **do**
    $P_a \leftarrow$ points $p \in \mathcal{P}$ whose closest center in A is a
  for $a \in A$ and $p \in P_a$ **do**
    $s(p) \leftarrow \frac{2\alpha \operatorname{dist}(p,A)^2}{\bar{c}} + \frac{4\alpha \sum_{p' \in P_a} \operatorname{dist}(p',A)^2}{|P_a|\bar{c}} + \frac{4|\mathcal{P}|}{|P_a|} + 1$
  for $p \in \mathcal{P}$ **do**
    $q(p) \leftarrow \frac{s(p)}{\sum_{p' \in \mathcal{P}} s(p')}$;
  $m \leftarrow O\Big(\frac{dk'^3 \log k'}{\epsilon^2}\Big)$
  $\mathcal{C} \leftarrow$ sample m weighted points from $\mathcal{P}$ where each point p has weight $\frac{1}{m \cdot q(p)}$ and is sampled with probability $q(p)$
  Return coreset $\mathcal{C}$

Step 2 We sample an $(\epsilon, \bar{k})$-coreset using the importance sampling scheme proposed in Algorithm 2. To bound the variance of our importance sampling scheme we sample each point with probability proportional to its sensitivity (Langberg & Schulman, 2010). The sensitivity $s(p)$ of a point $p \in \mathcal{P}$ is an upper bound on the maximum ratio between the cost contribution of the point and the average contribution of all points. We derive the necessary bounds based on the results for DP-Means++ and bound the required coreset size for the DP-Means clustering problem (see Theorem 3.3 and Supplementary Materials).



(a) DP-Means++ solution (b) Sampling probabilities (c) Coreset with weights (d) Solution

Figure 1. Coreset construction: (a) original dataset and the DP-Means++ approximate solution (red); (b) sampling probabilities (red signifies high probability); (c) resulting coreset and weights (red signifies a large weight); (d) cluster centers of the coreset solution (red).

(Step 3) To approximately solve the full problem, any DP-Means solver can finally be applied to the $(\epsilon, \bar{k})$-coreset if the solver can be extended to weighted data and to respect the known upper bound $\bar{k}$, i.e. it can be ensured that it only evaluates the DP-Means cost function for clusterings with less than $\bar{k}$ cluster centers. Both a brute-force approach based on solving K-Means for different values of k (explained in Section 4.2) and the DP-Means algorithm (Kulis & Jordan, 2012) satisfy this requirement.

3.3. Analysis

Our main contribution is stated in Theorem 3.3. It shows that the proposed method constructs valid $(\epsilon, \bar{k})$-coresets. This implies that, given an optimal solver for the DP-Means problem, an arbitrarily small approximation error can be obtained by solving on the coreset (Corollary 3.4).

Theorem 3.3. Let $0 < \epsilon < 1/4$, $\lambda > 0$ and let $\mathcal{P}$ be a set of n data points in $\mathbb{R}^d$. Let $\mathcal{C}$ be the weighted set returned by Algorithm 2 when applied to the result of Algorithm 1. Then the weighted set $\mathcal{C}$ is with constant probability an $(\epsilon, \hat{k})$-coreset with $\hat{k} = k'(16(\log_2 k' + 2) + 1)$ where k' is the number of centers returned by Algorithm 1. The result holds with probability at least $1 - \delta$ if Algorithm 1 is repeated $\log \frac{1}{\delta}$ times and $\mathcal{O}\left(\frac{dk'^3 \log k' + k'^2 \log \frac{1}{\delta}}{\epsilon^2}\right)$ points are sampled in Algorithm 2.

Proof sketch. The proof builds upon Theorem 4.1 and Theorem 4.4 of Feldman & Langberg (2011a). Firstly, we bound the sensitivities $s(p)$ for the DP-Means cost function using the (rough) approximation of the optimal solution obtained in DP-Means++. This allows us to derive the sampling probabilities $q(p)$. Secondly, we show that the total sensitivity is upper bounded by $\mathcal{O}(k')$ and that the dimension of the function space induced by the DP-Means cost function is upper bounded by $d(\bar{k} + 1)$ where $\bar{k}$ is the maximal number of cluster centers.

Corollary 3.4. Let $0 < \epsilon < 1/4$, $\lambda > 0$ and let $\mathcal{P}$ be a set of n data points in $\mathbb{R}^d$. Let $\mathcal{C}$ be the weighted set returned by Algorithm 2 when applied to the results of Algorithm 1. For any optimal solver Q mapping a set of data points to a set of cluster centers, we have $\text{cost}_{DP}(\mathcal{P}, Q(\mathcal{C})) \leq \frac{1+\epsilon}{1-\epsilon} \text{cost}_{DP}(\mathcal{P}, Q(\mathcal{P}))$

The number of points m to be sampled in Algorithm 2 depends on the size k' of the DP-Means++ solution. Our coresets are thus *data-dependent* and nonparametric since the size of the coreset scales with the complexity of the data. This stands in contrast to the fixed coreset sizes of existing constructions for parametric models. For theoretical completeness we provide comparable worst-case bounds on the required coreset size in Theorem 3.5. While there is an exponential dependency on d, the coreset size is sublinear in the number of data points n and only exhibits quadratic dependence on $1/\epsilon$.

Theorem 3.5. Let $0 < \epsilon < 1/4$, $\lambda > 0$ and let $\mathcal{P}$ be a set of n data points in $\mathbb{R}^d$. Let $\mathcal{C}$ be the weighted set returned by Algorithm 2 when applied to the result of Algorithm 1. Then the weighted set $\mathcal{C}$ has size at most $\tilde{\mathcal{O}}\left(d^{3d+2}k^{*3}(\log n)^3/\epsilon^2\right)$ where k^* is the optimal number of centers of the DP-Means clustering problem and the $\tilde{\mathcal{O}}(\cdot)$ notation subsumes any $\log k^*$, $\log d$ and $\log \log n$ terms.

Proof. The proof relies on bounding k' based on k^*. By setting ϵ to 1 in Theorem 3.2, we know that there exists a 1-coreset of size $m = \mathcal{O}(d^d k^* \log n)$. Consider all data points in such a coreset as a solution to the DP-Means problem. Its DP-Means cost on the same coreset is equal to $m\lambda$ as the quantization error is zero. The DP-Means error on the full data set is upper bounded by $2m\lambda$ due to the coreset property. The quantization error on the full data set is thus smaller than or equal to $m\lambda$ which in turn bounds the quantization error of the optimal K-Means++ solution for $k = m$. Since D^2-sampling provides us with a $\mathcal{O}(\log m)$ approximation, the stopping condition in Algorithm 1 is satisfied for m implying that k' is of $\mathcal{O}(d^d k^* \log n)$. □

4. Experimental results

In this section we validate our theoretical results and demonstrate the usefulness of our coreset construction. We use the following datasets:

- USGS (United States Geological Survey, 2010) — locations of 59'209 earthquakes between 1972 and 2010 mapped to 3D space using WGS 84.
- CSN (Faulkner et al., 2011) — 7GB of cellphone accelerometer data processed into 80'000 observations and 17 features.
- KDD (KDD Cup 2004, 2004) — 145'751 samples with 74 features measuring the match between a protein and a native sequence.
- MSYP (Bertin-Mahieux et al., 2011) — 90 features from 515'345 songs of the Million Song datasets used for predicting the year of songs.
- MNIST (LeCun et al., 1998) — 70'000 images of handwritten digits of size 28×28 pixels transformed using randomized PCA with whitening to 10 dimensions.

Our method works for instances of DP-Means with any value of λ; yet, for our experiments, we need to select a specific value. Choosing the "correct" value of λ for a dataset is a non-trivial task that is beyond the scope of this paper. For our purposes it is sufficient to use "reasonable" values of λ such that we can verify our theoretical results in a robust experimental setup. We want to ensure that our instances are neither degenerate (only one cluster in optimal solution) nor computationally infeasible (too many cluster centers in optimal solution). To ensure this, we solve K-Means for values of k between 5 and 200 to obtain a lower and upper bound on suitable values of λ. For each dataset we then select a value λ from this range (see Table 1) and roughly estimate the number of clusters $\tilde{k}$ in the optimal solution from the K-Means results. A similar approach was used by Kulis & Jordan (2012) where they first define the number of clusters k and then calculate λ based on a *farthest-first heuristic*.

4.1. Random evaluations

Both uniform subsampling and our coreset construction are instances of importance sampling and thus provide unbiased estimators of the cost function. The key difference is that coresets provide a bound on the variance leading to the coreset property as in Definition 3.1. In the first experiment we seek to validate this variance-reducing property.

We first obtain a random query Q by sampling $\tilde{k}$ points uniformly at random from the original dataset. We then construct a weighted subset $\mathcal{C}$ of the data using either uniform subsampling or our coreset construction method. We calculate the DP-Means cost on the full dataset $\mathrm{cost}_{DP}(\mathcal{P}, Q)$ and the weighted subset $\mathrm{cost}_{DP}(\mathcal{C}, Q)$. By repeating this procedure 500 times we obtain an unbiased estimator $\hat{\nu}^2$ of $\mathbb{E}\left[\nu^2\right]$, for both coresets and uniform subsampling where

$$\nu = (\mathrm{cost}_{DP}(\mathcal{C}, Q) - \mathrm{cost}_{DP}(\mathcal{P}, Q))/\,\mathrm{cost}_{DP}(\mathcal{P}, Q).$$

Table 1. Parametrization of λ for different datasets and corresponding estimated number of clusters in optimal solution $\tilde{k}$ (values of λ are not comparable as λ is not invariant of the data)

DATA SET	λ	$\tilde{k}$
USGS	1	160
KDD	10^9	60
CSN	10^3	60
MSYP	10^{10}	30
MNIST	10^3	70

Table 2. Normalized Shannon entropy of coreset sampling probabilities (a value of one signifies uniform sampling) and ratio between average variance ν_{cs} of coresets and average variance ν_{unif} of uniform subsampling

DATA SET	NORM. ENTROPY	ν_{cs}/ν_{unif}
USGS	0.97	0.33
KDD	0.94	0.01
CSN	0.85	0.01
MSYP	0.95	0.02
MNIST	0.99	0.75

Figure 2 shows the estimated variance for different subsample sizes on several datasets. As expected, the variance decreases for both coresets and the uniform subsample as the sample size increases. For all datasets except MNIST, we further observe that coresets exhibit a significantly lower variance than the uniform subsamples which confirms our theoretical results. Interestingly, for MNIST coresets and uniform subsample perform similarly. One might think that this is a failure of the coreset method. However, a closer look at Table 2 reveals that the calculated coreset sampling weights are almost uniform (the normalized entropy is close to one). It turns out that for this dataset uniform subsampling is comparable with coresets due to the balanced nature of the data. In contrast, for the other datasets the coreset construction leads to less uniform sampling weights. For those cases, coresets exhibit the expected variance-reducing property and outperform uniform subsampling by up to a factor of 100. This gap in variance widens, the more non-uniform the sampling weights are (see Table 2).

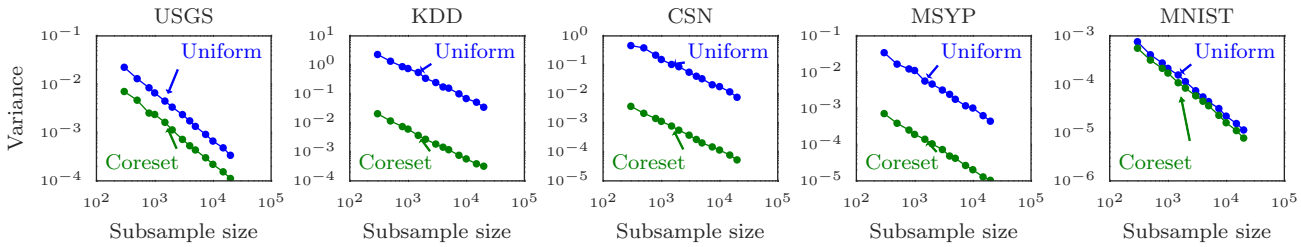


Figure 2. The variance of random query evaluations for coresets is lower or equal to the variance of uniform subsampling (averaged across 500 trials)

4.2. Solving DP-Means on the subsample

From the machine learning perspective, the ultimate interest in coresets is not the coreset property, but rather the solution to the original DP-Means clustering problem. In the following, we show that coresets indeed allow us to efficiently solve the clustering problem.

Again, we compute a weighted subsample of the data either through uniform subsampling or our coreset construction. We then solve the DP-Means clustering problem on the weighted subset and compute the DP-Means cost C_{ss} of this solution on the full dataset. We further solve the clustering problem on the full dataset and obtain the corresponding DP-Means cost C_{full}. We are finally interested in the relative error $\eta = (C_{ss} - C_{full})/C_{full}$. We use the following K-Means based procedure to solve the DP-Means clustering problem: We solve the K-Means clustering problem for different values of k chosen from a logarithmic grid[1] of 20 points between $2^{-1}\tilde{k}$ and $2^{2}\tilde{k}$ (see Table 1 for values of $\tilde{k}$). To solve one instance of the K-Means clustering we use a weighted version of K-Means++ and LLoyd's algorithm. We then compute the DP-Means cost for all solutions and choose the solution with the lowest error. The results in Table 3 demonstrate that this K-Means based solver significantly outperforms the original DP-Means algorithm (Kulis & Jordan, 2012) in terms of cost. The original DP-Means algorithm can easily get stuck in a local optimum with too few cluster centers, making it unsuitable for our experiments.

Figure 3 shows the relative DP-Means error η for both uniform subsampling and coresets compared to the full solution. Again, we observe that the cost for the weighted subsamples converges to the cost of the full solution as we increase the subsample size. As in the results for random evaluations, coresets outperform uniform subsampling except for MNIST where they perform similarly.

[1]The logarithmic grid leads to a very robust solver with results comparable across different runs for both coresets and uniform subsamples. In practice, if $\hat{k}$ is unknown, one can solve K-Means for exponentially increasing values of k until the DP-Means error decreases and then use binary search to find the final solution.

Table 3. DP-Means cost and number of clusters (#) for solutions based on K-Means grid solver and original DP-Means algorithm (Kulis & Jordan, 2012) as well as corresponding ratio (×) in cost

	GRID KM		ORIG. DP-MEANS		
DATASET	#	COST	#	COST	×
USGS	156	278.3	8	6634.7	23.8
KDD (10^9)	55	244.6	4	733.1	3.0
CSN (10^3)	58	167.5	16	379.8	2.3
MSYP (10^{12})	32	2.1	1	4.9	2.4
MNIST (10^3)	65	279.5	1	701.0	2.5

While coresets offer better performance, they come at a price. We need to invest a fixed time to compute the sampling probabilities for the coreset construction while uniform subsampling is essentially free. We thus investigate whether it makes sense to invest this time into the coreset construction instead of using it to solve the problem on a slightly larger uniform subsample of the data.

Figure 4 shows the DP-Means cost achieved in relation to the total time spent computing the subsample and solving the DP-Means problem. We observe that coresets offer a better time-quality tradeoff for USGS, KDD and CSN. For MSYP, the impact of construction time can be clearly seen. For a small time budget uniform subsampling is more efficient while coreset perform better for a larger time budget. For MNIST, coresets and uniform subsampling perform similarly; thus, coresets offer a slightly worse quality-time tradeoff due to the construction cost. Lucic et al. (2015) recently investigated tradeoffs in a more general setting and both empirically and theoretically demonstrated similar favorable time-quality tradeoffs for (parametric) coresets.

There are several reasons why coresets are preferable to uniform subsampling. While uniform subsampling is competitive on some datasets, it can fail arbitrarily badly on other datasets. On the other hand, coresets can be seen as offering insurance against such "bad" datasets as their theoretical guarantees hold on all datasets. The cost for this "insurance" is the fixed time we require for the construction

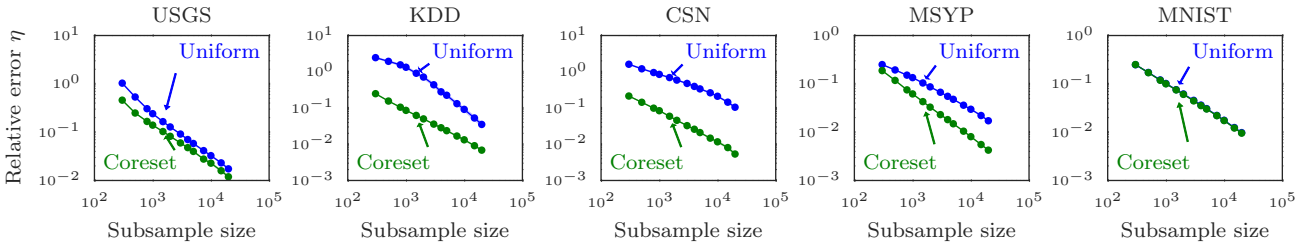


Figure 3. Relative DP-Means error using coresets is lower than the one using uniform subsamples for fixed subsample size (500 trials)

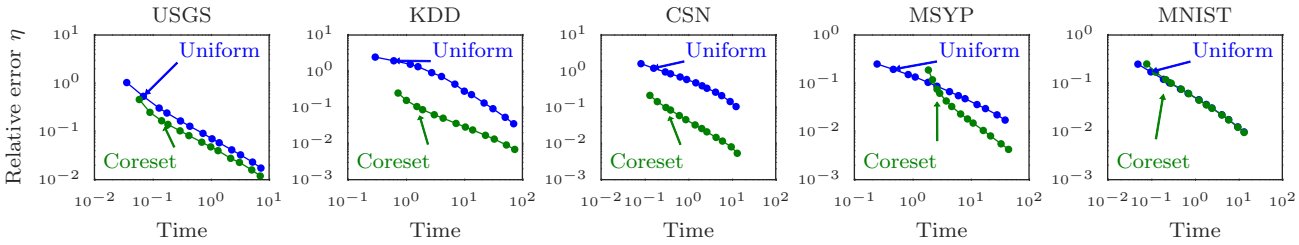


Figure 4. Relative DP-Means error for coresets and uniform subsamples compared to total time used (500 trials). In most settings, coresets require less time to achieve fixed error than uniform subsampling.

Table 4. Results for KDD (5000 subsamples)

	UNIFORM	CORESET	FULL
SAMPLING TIME (S)	0.01	0.49	-
SOLVER TIME (S)	13.33	13.49	635.07
TOTAL TIME (S)	13.33	13.98	635.07
SPEEDUP	47.6X	45.4X	1.0X
DP-MEANS COST (10^9)	299.54	250.36	244.57
RELATIVE ERROR η	22.5%	2.4%	0.0%

of the dataset which is almost linear in the number of data points, i.e., $\mathcal{O}(n \log n)$. Furthermore, in any practical case, we want to spend most time on solving the actual clustering problem by, e.g. choosing the subsample size reasonably large or running the solving algorithms several times with different initial seeds. This again reduces the impact of the construction time. For the case where the number of samples is prohibitively large to construct a coreset on the full dataset, a practical solution is to uniformly subsample the data to a size for which we can construct coresets and then to run the coreset generation. While having no theoretical guarantees with regards to the full clustering problem this generally still outperforms uniform subsampling.

The results in Table 4 for the KDD dataset and subsample size 5000 illustrate the practical relevance of coresets in speeding up inference for DP-Means. Instead of using the full data, coresets with size 3.43% of the full data allow us to achieve a speedup of 45.4 times with only 2.4% relative error. At the same time naive uniform subsampling leads to an relative error of 22.5% with a runtime comparable to that of coresets. The runtime of the sampling step for coresets, i.e., DP-Means++, is negligible requiring only 0.49 seconds or 3.5% of the total runtime of 13.98 seconds[2].

5. Conclusion

In this paper, we have demonstrated how coresets can be used to address large-scale nonparametric DP-Means clustering problems. We have shown the theoretical existence of coresets for queries of *arbitrary* size – a key theoretical challenge in applying coresets to Bayesian nonparametrics. We also showed that it is sufficient to bound the maximal size of queries dependent on the data and have provided a simple, practical algorithm providing such data dependent coresets. We also empirically demonstrated that this practical method allows us to scale inference in DP-Means by several orders of magnitude while retaining guarantees on the approximation error.

We believe that our results provide an important first step towards applying the technique of coresets to a more general class of Bayesian nonparametric models. We expect that – similar to recent developments in parametric models – it is possible to extend our techniques to more complex objective functions such as the log-likelihood of the Gaussian Dirichlet process mixture model or the HDP-Means problem (Kulis & Jordan, 2012). Coresets are a flexible tool as they only replace the full data set by a smaller representative set and thus can be combined with any method used to solve the original problem. As a result, we expect additional benefits by combining coresets with techniques for speeding up inference in Bayesian nonparametrics such as variational inference.

[2]All experiments were run on an Intel Xeon machine with 24 2.9GHz processors and 256GB RAM.

Acknowledgments

We would like to thank Sebastian Tschiatschek and the anonymous reviewers for their comments. This research was partially supported by ERC StG 307036 and the Zurich Information Security Center.

References

Antoniak, Charles E. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pp. 1152–1174, 1974.

Arthur, David and Vassilvitskii, Sergei. k-means++: The advantages of careful seeding. In *SODA*, pp. 1027–1035. SIAM, 2007.

Bertin-Mahieux, Thierry, Ellis, Daniel P.W., Whitman, Brian, and Lamere, Paul. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval*, 2011.

Blei, David M and Jordan, Michael I. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1): 121–143, 2006.

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

Escobar, Michael D and West, Mike. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

Faulkner, Matthew, Olson, Michael, Chandy, Rishi, Krause, Jonathan, Chandy, K Mani, and Krause, Andreas. The next big one: Detecting earthquakes and other rare events from community-based sensors. In *10th International Conference on Information Processing in Sensor Networks*, pp. 13–24. IEEE, 2011.

Feldman, Dan and Langberg, Michael. A unified framework for approximating and clustering data. In *STOC*, pp. 569–578. ACM, 2011a.

Feldman, Dan and Langberg, Michael. A unified framework for approximating and clustering data. Extended manuscript available at arXiv.org, 2011b.

Feldman, Dan, Monemizadeh, Morteza, and Sohler, Christian. A PTAS for k-means clustering based on weak coresets. In *SOCG*, pp. 11–18. ACM, 2007.

Feldman, Dan, Faulkner, Matthew, and Krause, Andreas. Scalable training of mixture models via coresets. In *NIPS*, pp. 2142–2150, 2011.

Har-Peled, Sariel and Kushal, Akash. Smaller coresets for k-median and k-means clustering. In *SOCG*, pp. 126–134. ACM, 2005.

Har-Peled, Sariel and Mazumdar, Soham. On coresets for k-means and k-median clustering. In *STOC*, pp. 291–300. ACM, 2004.

KDD Cup 2004. Protein Homology Dataset. Available at `http://osmot.cs.cornell.edu/kddcup/datasets.html`, 2004.

Kulis, Brian and Jordan, Michael I. Revisiting k-means: New algorithms via Bayesian nonparametrics. In *ICML*, pp. 513–520, 2012.

Langberg, Michael and Schulman, Leonard J. Universal ε-approximators for integrals. In *SODA*, pp. 598–607. SIAM, 2010.

LeCun, Yann, Bottou, Léon, Bengio, Yoshua, and Haffner, Patrick. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Lucic, Mario, Ohannessian, Mesrob I., Karbasi, Amin, and Krause, Andreas. Tradeoffs for space, time, data and risk in unsupervised learning. In *AISTATS*, 2015.

MacEachern, Steven N. Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3): 727–741, 1994.

MacEachern, Steven N, Clyde, Merlise, and Liu, Jun S. Sequential importance sampling for nonparametric bayes models: The next generation. *Canadian Journal of Statistics*, 27(2):251–267, 1999.

Neal, Radford M. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265, 2000.

Newton, Michael A and Zhang, Yunlei. A recursive algorithm for nonparametric analysis with missing data. *Biometrika*, 86(1):15–26, 1999.

Orbanz, Peter and Teh, Yee Whye. Bayesian nonparametric models. In *Encyclopedia of Machine Learning*, pp. 81–89. Springer, 2010.

United States Geological Survey. Global earthquakes (1.1.1972-19.3.2010). Retrieved from the mldata.org repository `https://mldata.org/repository/data/viewslug/global-earthquakes/`, 2010.

6. Appendix

We first prove the theoretical existence result of Theorem 3.2, then provide an analysis of the DP-Means++ algorithm and finally show that the importance sampling scheme in Algorithm 2 produces valid coresets.

6.1. Proof of Theorem 3.2

Theorem 3.2. Let $0 < \epsilon \leq 1$ and let $\mathcal{P}$ be a set of n points in $\mathbb{R}^d$. Then there exists an ϵ-coreset for the DP-Means clustering of $\mathcal{P}$ with size $\mathcal{O}\Big(\frac{d^d k^* \log n}{\epsilon^d}\Big)$ where k^* is the optimal number of centers.

Proof. To show existence we provide a coreset construction based on the exponential grid approach by Har-Peled & Mazumdar (2004). Let the k^*-sized set $A = \{a_1, a_2, \ldots, a_{k^*}\}$ be an optimal solution to the DP-Means clustering problem. We partition the dataset $\mathcal{P}$ into subsets $\{P_i\}_{i=1,2,\ldots,k^*}$ by assigning each point in $\mathcal{P}$ to the closest center in A and we define $R = \sqrt{\frac{\text{cost}_{DP}(\mathcal{P},A)}{n}}$.

As in Har-Peled & Mazumdar (2004), we build an exponential grid around the centroids $A = \{a_1, a_2, \ldots, a_{k^*}\}$. Let $M = \lceil \log_2(n)/2 \rceil$ and define the axis-parallel squares $Q_{i,j}$ with side length $R2^{j+1}$ centered at a_i for $j = 0, 1, \ldots, M$ and $i = 1, 2, \ldots, k^*$. Let $V_{i,0} = Q_{i,0}$ for $i = 1, 2, \ldots, k^*$ and let $V_{i,j} = Q_{i,j} \setminus Q_{i,j-1}$ for $j = 1, 2, \ldots, M$ and $i = 1, 2, \ldots, k^*$. Now we partition each $V_{i,j}$ into a grid with side length $r_j = \frac{\epsilon}{10d} R2^{j-1}$ and denote the resulting exponential grid by G_i. Finally, we assign all points in P_i to their grid cell in G_i and pick an arbitrary representative for each grid cell with weight equal to the number of points in the cell. The coreset $\mathcal{C}$ is formed by the union of all representative points.

We first show by contradiction that all points in $\mathcal{P}$ fall within a cell of the resulting exponential grid $G = \cup_i G_i$. Suppose $p \in \mathcal{P}$ is not in a grid cell. Then we have

$$\text{dist}(p, A) > R2^M \geq R\sqrt{n}.$$

This implies

$$\text{dist}(p, A)^2 > nR^2 = \sum_{p' \in \mathcal{P}} \text{dist}(p', A)^2$$

which leads to the contradiction $\text{dist}(p, A)^2 > \text{dist}(p, A)^2$.

In order for $\mathcal{C}$ to be a valid ϵ-coreset, we need to show that for any query $B \subset \mathbb{R}^d$

$$|\text{cost}_{DP}(\mathcal{P}, B) - \text{cost}_{DP}(\mathcal{C}, B)| \leq \epsilon \, \text{cost}_{DP}(\mathcal{P}, B) \quad (1)$$

For any point $p \in \mathcal{P}$ we denote by p' the corresponding image (the closest point) in the coreset $\mathcal{C}$ and by $\text{dist}(p, p')$ the distance between p and p'. Using the triangle inequality, we deduce that

$$\begin{aligned}
\mathcal{E} &= |\text{cost}_{DP}(\mathcal{P}, B) - \text{cost}_{DP}(\mathcal{C}, B)| \\
&= \left| \sum_{p \in \mathcal{P}} \text{dist}(p, B)^2 - \sum_{(p,w) \in \mathcal{C}} w \, \text{dist}(p', B)^2) \right| \\
&\leq \sum_{p \in \mathcal{P}} \big| \text{dist}(p, B)^2 - \text{dist}(p', B)^2 \big| \\
&\leq \sum_{p \in \mathcal{P}} |(\text{dist}(p, B) - \text{dist}(p', B))(\text{dist}(p, B) + \text{dist}(p', B))| \\
&\leq \sum_{p \in \mathcal{P}} \text{dist}(p, p')(2 \, \text{dist}(p, B) + \text{dist}(p, p')).
\end{aligned}$$

To bound the total error, we partition the set $\mathcal{P}$ into the three subsets P_R, P_A and P_B, and bound the error separately. For the set $P_R = \{p \in \mathcal{P} \mid \text{dist}(p, B) \leq R \wedge \text{dist}(p, A) \leq R\}$ we have $\text{dist}(p, p') \leq \frac{\epsilon}{10} R$ and $\text{dist}(p, B) \leq R$. Hence

$$\begin{aligned}
\mathcal{E}_R &= \sum_{p \in P_R} \text{dist}(p, p')(2 \, \text{dist}(p, B) + \text{dist}(p, p')) \\
&\leq \sum_{p \in P_R} \frac{\epsilon}{10} R \Big(2R + \frac{\epsilon}{10} R \Big) \\
&\leq \sum_{p \in P_R} \frac{\epsilon}{3} R^2 \leq \frac{\epsilon}{3} \sum_{p \in \mathcal{P}} \text{dist}(p, A)^2.
\end{aligned}$$

For the set $P_A = \{p \in \mathcal{P} \setminus P_R \mid \text{dist}(p, B) \leq \text{dist}(p, A)\}$ we use that $\text{dist}(p, p') \leq \frac{\epsilon}{10} \text{dist}(p, A)$ and $\text{dist}(p, A) \leq R_\lambda$. Thus

$$\begin{aligned}
\mathcal{E}_A &= \sum_{p \in P_A} \text{dist}(p, p')(2 \, \text{dist}(p, B) + \text{dist}(p, p')) \\
&\leq \sum_{p \in P_A} \frac{\epsilon}{10} \Big(2 + \frac{\epsilon}{10} \Big) \text{dist}(p, A)^2 \\
&\leq \frac{\epsilon}{3} \sum_{p \in P_A} \text{dist}(p, A)^2.
\end{aligned}$$

For the remaining set $P_B = \mathcal{P} \setminus (P_R \cup P_A)$ we use that $\text{dist}(p, p') \leq \frac{\epsilon}{10} \text{dist}(p, A) \leq \frac{\epsilon}{10} \text{dist}(p, B)$ since $\text{dist}(p, B) > R$ and $\text{dist}(p, B) > \text{dist}(p, A)$. Hence

$$\begin{aligned}
\mathcal{E}_B &= \sum_{p \in P_B} \text{dist}(p, p')(2 \, \text{dist}(p, B) + \text{dist}(p, p')) \\
&\leq \sum_{p \in P_B} \frac{\epsilon}{10} \Big(2 + \frac{\epsilon}{10} \Big) \text{dist}(p, B)^2 \\
&\leq \frac{\epsilon}{3} \sum_{p \in P_B} \text{dist}(p, B)^2.
\end{aligned}$$

The optimality of A allows us to bound $\mathcal{E}$ and thus show

that (1) holds for all finite $B \subset \mathbb{R}^d$

$$\begin{aligned}\mathcal{E} &\leq \mathcal{E}_R + \mathcal{E}_A + \mathcal{E}_B \\ &\leq \frac{2\epsilon}{3} \sum_{p \in P_A} \operatorname{dist}(p, A)^2 + \frac{\epsilon}{3} \sum_{p \in P_B} \operatorname{dist}(p, B)^2 \\ &\leq \epsilon \operatorname{cost}_{DP}(\mathcal{P}, B).\end{aligned}$$

To complete the proof note that each of the $V_{i,j}$ contains at most $\mathcal{O}(d^d \epsilon^{-d})$ cells which implies that the resulting set $\mathcal{C}$ is of size $\mathcal{O}\Big(\frac{d^d k^* \log(n)}{\epsilon^d}\Big)$. □

6.2. Analysis of DP-Means++

We first show that DP-Means++ is $\mathcal{O}(\log k')$-competitive to the optimal solution and then prove that the optimal number of centers k^* is bounded by $\mathcal{O}(k' \log k')$.

Lemma 6.1. Let $\lambda > 0$. Suppose the set A was sampled using Algorithm 1. Then with probability at least $1/2$ the set A is a $\mathcal{O}(\log k')$-competitive solution to the DP-Means optimization problem with $k' = |A|$, i.e.

$$\operatorname{cost}_{DP}(\mathcal{P}, A) \leq (16 \log_2 k' + 34) \operatorname{OPT}_{DP}(\mathcal{P}, \lambda). \quad (2)$$

The probability can be boosted to $1 - \delta$ by repeating the sampling $\log(1/\delta)$ times and picking the solution with the lowest DP-Means cost.

Proof. We first recall a main result on D^2-sampling. Denote by Δ_k the quantization error after k rounds of D^2-sampling and by Δ_k^* the optimal quantization error for K-Means with k cluster centers. Arthur & Vassilvitskii (2007) show that in expectation D^2-sampling is $\mathcal{O}(\log k)$-competitive, i.e.

$$\mathbb{E}[\Delta_k] \leq 8(\log_2 k + 2)\Delta_k^* \quad (3)$$

Algorithm 1 is an instance of D^2-sampling that stops after k' iterations such that for k'

$$\frac{\Delta_{k'}}{16(\log_2 k' + 2)} < k'\lambda \quad (4)$$

while for $k' - 1$

$$\frac{\Delta_{k'-1}}{16(\log_2(k' - 1) + 2)} \geq (k' - 1)\lambda. \quad (5)$$

Let k^* denote the number of centers in the optimal solution of the DP-Means problem. We show (2) by considering two cases for k^*.

Case $k^* < k'$. In this case the quantization error of the optimal clustering with $k' - 1$ centers bounds the optimal DP-Means cost from below, i.e.

$$\operatorname{OPT}_{DP}(\mathcal{P}, \lambda) \geq \Delta_{k'-1}^*.$$

Applying Markov's inequality to (3), we have with probability at least $1/2$

$$\Delta_{k'-1} \leq 16(\log_2[k' - 1] + 2)\Delta_{k'-1}^*.$$

This result in combination with (5) implies

$$\Delta_{k'-1}^* \geq (k' - 1)\lambda.$$

With probability at least $1/2$, we thus have

$$\begin{aligned}\operatorname{cost}_{DP}(\mathcal{P}, A) &= \Delta_{k'} + k'\lambda \\ &\leq \Delta_{k'-1}^* + 2(k' - 1)\lambda \\ &\leq 16(\log_2[k' - 1] + 2)\Delta_{k'-1}^* + 2\Delta_{k'-1}^* \\ &\leq (16 \log_2 k' + 34) \operatorname{OPT}_{DP}(\mathcal{P}, \lambda).\end{aligned}$$

Case $k^* \geq k'$. In this case, the optimal DP-Means cost is bounded from below by

$$\operatorname{OPT}_{DP}(\mathcal{P}, \lambda) \geq k'\lambda.$$

This in combination with (4) is sufficient to show

$$\begin{aligned}\operatorname{cost}_{DP}(\mathcal{P}, A) &= \Delta_{k'} + k'\lambda \\ &\leq 16(\log_2 k' + 2)k'\lambda + k'\lambda \\ &\leq (16 \log_2 k' + 33) \operatorname{OPT}_{DP}(\mathcal{P}, \lambda)\end{aligned}$$

which concludes the proof. □

Note that Algorithm 1 has an additional property: after running it, we already obtain an upper bound on the optimal number of clusters k^*.

Lemma 6.2. Let $\lambda > 0$. Denote by k^* the optimal number of centers for the DP-Means clustering problem and suppose k' points are sampled using Algorithm 1 with parameter λ. Then

$$k^* \leq k'(16(\log_2 k' + 2) + 1)$$

Proof. Denote by $\Delta_{k'}$ the quantization error after k' iterations. The result then follows from the fact that $k^* \leq k' + \Delta_{k'}/\lambda$ and (4). □

6.3. Analysis of importance sampling scheme

Our coreset construction builds upon the framework of Feldman & Langberg (2011a) in which it is shown that importance sampling with probabilities proportional to the "sensitivity" of each point leads to valid coresets. The number of required points depends on the paramter ϵ, the bound on the total sensitivity S and the "combinatorial complexity" induced by the query space $\mathcal{Q}$ and the cost function $f_Q(\cdot)$. This notion is formalized in Definition 6.3 and the main theorem is presented in Theorem 6.4.

Definition 6.3 (Feldman & Langberg (2011a))**.** Let $\mathcal{P}$ be a finite set and denote by $f_Q(p)$ a cost function from $\mathcal{Q}\times\mathcal{P}$ to $[0,\infty)$. Define the set of functions $F = \{f_Q(p) \mid p \in \mathcal{P}\}$ from the set $\mathcal{Q}$ to $[0,\infty)$. The dimension $\dim(F)$ of F is the minimum integer d such that

$$\forall S \subseteq F : |S \cap \mathbf{ranges}(F)| \leq |S|^d$$

where $\mathbf{ranges}(F) = \{\mathbf{range}(Q,r) \mid Q \in \mathcal{Q}, r \geq 0\}$ and $\mathbf{range}(Q,r) = \{f \in F \mid f(Q) \leq r\}$ for every $Q \in \mathcal{Q}$ and $r \geq 0$.

Theorem 6.4 (Feldman & Langberg (2011a))**.** Let $0 < \epsilon < 1/4$. Let $\mathcal{P}$ be a finite set and denote by $f_Q(p)$ a cost function from $\mathcal{Q} \times \mathcal{P}$ to $[0,\infty)$. Define the set of functions $F = \{f_Q(p) \mid p \in \mathcal{P}\}$ from the set $\mathcal{Q}$ to $[0,\infty)$. Let $s : \mathcal{P} \to \mathbb{N} \setminus \{0\}$ be a function such that

$$s(p) \geq \max_{Q\in\mathcal{Q}} \frac{f_Q(p)}{\bar{f}_Q}$$

and let $S = \sum_{p\in\mathcal{P}} s(p)/n$. For each $p \in P$, let $g_p : \mathcal{Q} \to [0,\infty)$ be defined as $g_p(Q) = f_Q(p)/s(p)$. Let G_p consist of s_p copies of g_p, and let C be a random sample of

$$t = \frac{\dim(F)\Big(\sum_{p\in\mathcal{P}} s(p)\Big)^2}{(\epsilon n)^2} = \frac{\dim(F)S^2}{\epsilon^2}$$

functions from the set $G = \bigcup_{p\in\mathcal{P}} G_p$. Then for every $Q \in \mathcal{Q}$

$$\left|\sum_{p\in\mathcal{P}} f_Q(p) - \sum_{c\in C} g_c(Q)\right| \leq \epsilon \sum_{p\in\mathcal{P}} f_Q(p).$$

We use these results to show that our sampling scheme produces valid coresets. We first derive sampling probabilities for the DP-Means problem and then bound the total sensitivity S as well as $\dim(F)$ to obtain the coreset size.

Lemma 6.5. Let the set A be an α-competitive solution to the DP-Means clustering problem with β centers. Let $\mathcal{P}$ be the point set of size n. For each point $p \in \mathcal{P}$ denote by a_i the closest center in the set A and by P_i the containing Voronoi cell induced by A on $\mathcal{P}$. Then for each point $p \in \mathcal{P}$ the sensitivity $\sigma(p) = \max_{Q\subset\mathbb{R}^d} f_Q(p)/\bar{f}_Q$ is bounded by

$$\begin{aligned} s(p) =& \frac{2\alpha\,\mathrm{dist}(p,A)^2}{\mathrm{cost}_{DP}(\mathcal{P},A)/n} + \frac{4\alpha\sum_{p'\in P_i}\mathrm{dist}(p',A)^2}{|P_i|\,\mathrm{cost}_{DP}(\mathcal{P},A)/n} \\ &+ \frac{4n}{|P_i|} + 1 \end{aligned}$$

and the total sensitivity $\mathfrak{S} = \sum_{p\in\mathcal{P}} \sigma(p)/n$ is bounded by

$$S = 6\alpha + 4\beta + 1.$$

Proof. Using the triangle inequality and the fact that $(a+b)^2 \leq 2a^2 + 2b^2$ for non-negative a and b, we have for any query $W \subset \mathbb{R}^d$, any center $a_i \in A$ and any point $p \in P_i$ in the center's Voronoi region P_i

$$\mathrm{dist}(a_i,Q)^2 \leq 2\,\mathrm{dist}(p,a_i)^2 + 2\,\mathrm{dist}(p,Q)^2.$$

For each center $a_i \in A$ we may sum all points in the Voronoi region P_i to get

$$\mathrm{dist}(a_i,Q)^2 \leq \frac{2}{|P_i|}\left[\sum_{p\in P_i}\mathrm{dist}(p,a_i)^2 + \sum_{p\in P_i}\mathrm{dist}(p,Q)^2\right].$$

For any query Q we have

$$\begin{aligned} \frac{f_Q(p)}{\bar{f}_Q} \leq& \frac{\mathrm{dist}(p,Q)^2 + |Q|\lambda/n}{\sum_{p'\in\mathcal{P}}\mathrm{dist}(p',Q)^2/n + |Q|\lambda/n} \\ \leq& \frac{2\,\mathrm{dist}(p,a_i)^2}{\mathrm{OPT}_{DP}(\mathcal{P},\lambda)/n} \\ &+ \frac{2\,\mathrm{dist}(a_i,Q)^2}{\sum_{p'\in\mathcal{P}}\mathrm{dist}(p',Q)^2/n + |Q|\lambda/n} + 1 \\ \leq& \frac{2\,\mathrm{dist}(p,a_i)^2}{\mathrm{OPT}_{DP}(\mathcal{P},\lambda)/n} + \frac{4\sum_{p'\in P_i}\mathrm{dist}(p',a_i)^2}{|P_i|\,\mathrm{OPT}_{DP}(\mathcal{P},\lambda)/n} \\ &+ \frac{4\sum_{p\in P_i}\mathrm{dist}(p',Q)^2}{|P_i|\sum_{p'\in\mathcal{P}}\mathrm{dist}(p',Q)^2/n} + 1 \\ \leq& \frac{2\,\mathrm{dist}(p,a_i)^2}{\mathrm{OPT}_{DP}(\mathcal{P},\lambda)/n} + \frac{4\sum_{p'\in P_i}\mathrm{dist}(p',a_i)^2}{|P_i|\,\mathrm{OPT}_{DP}(\mathcal{P},\lambda)/n} \\ &+ \frac{4n}{|P_i|} + 1 \\ \leq& \frac{2\alpha\,\mathrm{dist}(p,A)^2}{\mathrm{cost}_{DP}(\mathcal{P},A)/n} + \frac{4\alpha\sum_{p'\in P_i}\mathrm{dist}(p',A)^2}{|P_i|\,\mathrm{cost}_{DP}(\mathcal{P},A)/n} \\ &+ \frac{4n}{|P_i|} + 1. \end{aligned}$$

The last expression is independent of Q and thus a valid upper bound on $\sigma(p)$. Furthermore, the total sensitivity $\mathfrak{S}$ is bounded by S since

$$\begin{aligned} \mathfrak{S} =& \frac{1}{n}\sum_{p\in\mathcal{P}}\sigma(p) \\ \leq& \sum_{p\in\mathcal{P}}\left[\frac{2\alpha\,\mathrm{dist}(p,A)^2}{\mathrm{cost}_{DP}(\mathcal{P},A)} + \frac{4\alpha\sum_{p'\in P_i}\mathrm{dist}(p',A)^2}{|P_i|\,\mathrm{cost}_{DP}(\mathcal{P},A)}\right. \\ &\left. + \frac{4}{|P_i|}\right] + 1 \\ =& \frac{6\alpha\sum_{p\in\mathcal{P}}\mathrm{dist}(p,A)^2}{\mathrm{cost}_{DP}(\mathcal{P},A)} + 4\beta + 1 \\ \leq& 6\alpha + 4\beta + 1 = S. \end{aligned}$$

□

Lemma 6.6. Let $\mathcal{Q}$ be the set of all non-empty subsets of $\mathbb{R}^d$ with at most k elements. Let $\mathcal{P}$ be a finite set and denote by $f_Q(p)$ the DP-Means cost function from $\mathcal{Q} \times \mathcal{P}$ to $[0, \infty)$. Define the set of functions $F = \{f_Q(p) \mid p \in \mathcal{P}\}$ from the set $\mathcal{Q}$ to $[0, \infty)$. Then it holds that

$$\dim(F) = d(k+1).$$

Proof. We need to show that

$$\forall S \subseteq F : |S \cap \mathbf{ranges}(F)| \leq |S|^{d(k+1)}$$

where $\mathbf{ranges}(F) = \{\mathbf{range}(Q, r) \mid Q \in \mathcal{Q}, r \geq 0\}$ and $\mathbf{range}(Q, r) = \{f \in F \mid f(Q) \leq r\}$ for every $Q \in \mathcal{Q}$ and $r \geq 0$.

This is equivalent to showing $\forall S \subseteq F$

$$\sum_{i=1,\ldots,k} |S \cap \{\mathbf{range}(Q, r) \mid Q \in \mathcal{Q}_i, r \geq 0\}| \leq |S|^{kd+1}$$

where $\mathcal{Q}_i$ is the set of all subsets of $\mathbb{R}^d$ with i elements.

We recall the DP-Means cost function $f_Q(p) = \mathrm{dist}(p, Q)^2 + \frac{|Q|\lambda}{n}$ and easily see that for each query Q and $r \geq 0$ the range $\mathbf{range}(Q, r)$ for the DP-Means problem is the same as the equivalently defined range $\mathbf{range}_{KM}(Q, r - \frac{|Q|\lambda}{n})$ of the K-Means clustering problem with $|Q|$ centers since the DP-Means penalization term $\frac{|Q|\lambda}{n}$ is constant. We hence use the result from Feldman & Langberg (2011b) (see proof of Theorem 15.7) that in d-dimensional Euclidean space K-Means clustering with i centers has dimensionality id which implies for $i = 1, 2, \ldots, k$

$$\forall S \subseteq F : |S \cap \{\mathbf{range}(Q, r) \mid Q \in \mathcal{Q}_i, r \geq 0\}| \leq |S|^{id}.$$

This gives us the required result

$$\forall S \subseteq F : |S \cap \mathbf{ranges}(F)| \leq \sum_{i=1,2,\ldots,k} |S|^{id} \leq |S|^{d(k+1)}.$$

□

Now we have all the prerequisites to prove our main results, Theorem 3.3 and Corollary 3.4.

Theorem 3.3. Let $0 < \epsilon < 1/4$, $\lambda > 0$ and let $\mathcal{P}$ be a set of n data points in $\mathbb{R}^d$. Let $\mathcal{C}$ be the weighted set returned by Algorithm 2 when applied to the results of Algorithm 1. Then the weighted set $\mathcal{C}$ is with constant probability an $(\epsilon, \hat{k})$-coreset with $\hat{k} = k'(16(\log_2 k' + 2) + 1)$ where k' is the number of centers returned by Algorithm 1. The result holds with probability at least $1 - \delta$ if Algorithm 1 is repeated $\log \frac{1}{\delta}$ times and $\mathcal{O}\Big(\frac{dk'^3 \log k' + k'^2 \log \frac{1}{\delta}}{\epsilon^2}\Big)$ points are sampled in Algorithm 2.

Proof. Apply Lemma 6.1, Lemma 6.5 and Lemma 6.6 to Theorem 6.4. The results can be extended to hold with arbitrary probability $1 - \delta$ by Theorem 4.4 of Feldman & Langberg (2011a). □

Corollary 3.4. Let $0 < \epsilon < 1/4$, $\lambda > 0$ and let $\mathcal{P}$ be a set of n data points in $\mathbb{R}^d$. Let $\mathcal{C}$ be the weighted set returned by Algorithm 2 when applied to the results of Algorithm 1. For any optimal solver Q mapping a set of data points to a set of cluster centers, we have $\mathrm{cost}_{DP}(\mathcal{P}, Q(\mathcal{C})) \leq \frac{1+\epsilon}{1-\epsilon} \mathrm{cost}_{DP}(\mathcal{P}, Q(\mathcal{P}))$

Proof. Since the solver Q is optimal, we have

$$\mathrm{cost}_{DP}(\mathcal{C}, Q(\mathcal{C})) \leq \mathrm{cost}_{DP}(\mathcal{C}, Q(\mathcal{P})).$$

By Theorem 3.3 we have both

$$\mathrm{cost}_{DP}(\mathcal{C}, Q(\mathcal{C})) \geq (1-\epsilon)\,\mathrm{cost}_{DP}(\mathcal{P}, Q(\mathcal{C}))$$
$$\mathrm{cost}_{DP}(\mathcal{C}, Q(\mathcal{P})) \leq (1+\epsilon)\,\mathrm{cost}_{DP}(\mathcal{P}, Q(\mathcal{P})).$$

This leads to the desired result, i.e.

$$\mathrm{cost}_{DP}(\mathcal{P}, Q(\mathcal{C})) \leq \frac{1+\epsilon}{1-\epsilon} \mathrm{cost}_{DP}(\mathcal{P}, Q(\mathcal{P})).$$

□