# Wasserstein Coresets for Lipschitz Costs

**Sebastian Claici**
CSAIL
MIT
Cambridge, MA 02139
sclaici@mit.edu

**Justin Solomon**
CSAIL
MIT
Cambridge, MA 02139
jsolomon@mit.edu

## Abstract

Sparsification is becoming more and more relevant with the proliferation of huge data sets. Coresets are a principled way to construct representative weighted subsets of a data set that have matching performance with the full data set for specific problems. However, coreset language neglects the nature of the underlying data distribution, which is often continuous. In this paper, we address this oversight by introducing a notion of *measure coresets* that generalizes coreset language to arbitrary probability measures. Our definition reveals a surprising connection to *optimal transport* theory which we leverage to design a coreset for problems with Lipschitz costs. We validate our construction on support vector machine (SVM) training, $k$-means clustering, $k$-median clustering, and linear regression and show that we are competitive with previous coreset constructions.

## 1 Introduction

Data sets with hundreds of millions of examples are becoming the norm in machine learning, whether for Bayesian inference, clustering, or regression. The complexity of algorithms for these tasks typically scales in the size of the data set, making it difficult to employ the entire input effectively. Several techniques attempt to overcome this challenge for scalable machine learning, from streaming optimization to subsampling; these often reduce computational complexity on large data sets but are accompanied with weak theoretical guarantees.

Originally proposed for computational geometry (Agarwal et al., 2005), *coresets* recently have been introduced as a principled means of reducing input sizes for machine learning algorithms. Intuitively, a coreset of a data set is a "representative" subsample on which a given machine learning algorithm is guaranteed to produce similar output. Coresets have been applied successfully to learning tasks including clustering (Bachem et al., 2018), classification via SVMs (Tsang et al., 2005), neural network compression (Baykal et al., 2018), and Bayesian inference (Huggins et al., 2016).

Coreset computation is typically posed as a discrete problem: Given a fixed data set and learning algorithm, how can we construct a smaller data set from which the learning algorithm will have similar performance? This posing of the problem is compatible with descriptions of coresets in computational geometry but neglects a key theme in machine learning: a data set is nothing more than an empirical sample from an underlying data distribution—the latter being the key to describing a learning task. That is, we typically do not need a coreset for a specific data set but rather for the distribution from which it was drawn.

To address this gap, in this paper we extend the definition of a coreset to (possibly smooth) data distributions, where the classical notion of a coreset is recovered when the distribution is composed of a finite set of deltas (an empirical distribution). Beyond broadening the definition, we show that a sufficient condition for extracting a coreset can be written in the language of *optimal transport*. For Lipschitz cost functions, our approach defines a new framework for coreset analysis and construction.

Preprint. Work in progress.

We apply our constructions to extract coresets for classification and clustering, with performance competitive with that of previous techniques and broader generality. We rely on a recently proposed stochastic algorithm for approximating distributions in the Wasserstein metric (Claici et al., 2018). In brief, we construct a discrete distribution supported on $n$ points that minimizes distance to the original distribution in the Wasserstein metric. We can show that several methods for classification and clustering have costs that are upper-bounded by the Wasserstein distance between the learned and true data distributions. This leads to fairly compact coresets that perform well in practice.

**Contributions.** We give a practical, parallelizable algorithm for coreset construction for a variety of problems, backed by an interpretation of coresets in continuous probabilistic language. Our construction reveals a surprising tie-in with *optimal transport* theory. Our coresets are among few in machine learning that do not rely on importance sampling. We prove bounds on the size of the coreset that are independent of data set size and generalize to any machine learning problem whose cost is separable and Lipschitz. Finally, we compare with state-of-the-art on SVM binary classification, linear regression, $k$-means clustering, and $k$-median clustering, and show competitive performance.

### 1.1 Related work

Our work lies at the intersection of two seemingly unrelated lines of research, joining the probabilistic language of optimal transport research with the discrete setting of data compression via coresets.

**Coresets.** Initially introduced for problems in computational geometry (Agarwal et al., 2005), coresets have found their way to machine learning research via importance sampling (Langberg & Schulman, 2010). Coreset applications are varied, and generic frameworks exist for constructing them for almost any problem (Feldman & Langberg, 2011). Among the most relevant and recent applications are $k$-means and $k$-median clustering (Har-Peled & Mazumdar, 2004; Arthur & Vassilvitskii, 2007; Feldman et al., 2013; Bachem et al., 2018), Bayesian inference (Campbell & Broderick, 2018; Huggins et al., 2016), support vector machine training (Tsang et al., 2005), and neural network compression (Baykal et al., 2018). While coreset language is discrete, the sensitivity-based approach that importance sampling coresets use currently was introduced in a continuous setting for approximating expectations of functions under measures that are absolutely continuous with respect to the Lebesgue measure (Langberg & Schulman, 2010). For more information, see the introduction by Bachem et al. (2018) and the survey paper by Munteanu & Schwiegelshohn (2018).

**Optimal Transport** is a relative newcomer to machine learning. Sparked by advances in entropically-regularized transport (Cuturi, 2013), an influx of applications of optimal transport to machine learning have appeared, from supervised learning (Schmitz et al., 2017; Carrière et al., 2017), to Bayesian inference (Staib et al., 2017; Srivastava et al., 2015) and neural network training (Arjovsky et al., 2017; Montavon et al., 2016; Genevay et al., 2018). Details can be found in several recent surveys (Solomon, 2018; Peyré & Cuturi, 2018; Lévy & Schwindt, 2017).

Our approach is inspired by semi-discrete methods that compute transport from a continuous measure to a discrete one by leveraging power diagrams (Aurenhammer, 1987). Efficient algorithms that use computational geometry tools to perform gradient iterations to solve the Kantorovich dual problem have been introduced for 2D (Mérigot, 2011) and 3D (Lévy, 2015). Closer to our method are the algorithms by De Goes et al. (2012) and Claici et al. (2018) that solve a non-convex problem for the support of a discrete uniform measure that minimizes transport cost to an input image (De Goes et al., 2012) or the barycenter of the input distributions (Claici et al., 2018).

## 2 Optimal Transport and Coresets

### 2.1 Optimal Transport

Optimal transport measures distances between probability distributions in a geometric fashion. Given a metric space $(\mathcal{X}, d)$ and measures $\mu_1, \mu_2 \in \mathcal{P}(\mathcal{X})$, we define the $p$-Wasserstein (transport) cost:

$$W_p(\mu_1, \mu_2) = \left( \inf_{\gamma \in \Gamma(\mu_1, \mu_2)} \int_{\mathcal{X} \times \mathcal{X}} d(\mathbf{x}, \mathbf{y})^p \, \mathrm{d}\gamma(\mathbf{x}, \mathbf{y}) \right)^{1/p},$$

where $\Gamma(\mu_1, \mu_2) \subset \mathcal{P}(\mathcal{X} \times \mathcal{X})$ is the set of measure couplings between $\mu_1$ and $\mu_2$:

$$\Gamma(\mu_1, \mu_2) = \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) : (\pi_x)_{\#}\gamma = \mu_1, (\pi_y)_{\#}\gamma = \mu_2\}.$$

### 2.2 Coresets

Let $\mathcal{F}$ be a family of functions from $\mathbb{R}^d$ to $\mathbb{R}$, and $X = \{\mathbf{x}_i\}_{i=1}^N$ a subset of $\mathbb{R}^d$ with weighting function $\mu_X : X \to \mathbb{R}_+$. We define the cost of $X$ for a given $f \in \mathcal{F}$ as

$$\text{cost}(X, f) = \sum_{\mathbf{x} \in X} \mu_X(\mathbf{x}) f(\mathbf{x}). \tag{1}$$

A coreset is a weighted subset of the initial data that yields similar costs to the full data set at a fraction of the size. Formally,

**Definition 1.** *The pair $(C, \mu_C)$ is an $\varepsilon$-coreset for the cost function* cost *and the function family $\mathcal{F}$ if $C \subseteq X$ and $|\text{cost}(X, f) - \text{cost}(C, f)| \leq \varepsilon \cdot \text{cost}(X, f)$ for all $f \in \mathcal{F}$.*

A coreset always exists as we can take $C = X$ and $\mu_C = \mu_X$ to satisfy the inequality.

Typically, coresets are problem-specific, for example applied to $k$-means clustering or SVM classification. In such cases, the function family (or query set) reflects the cost of the original problem. For example, for $k$-means clustering the family $\mathcal{F}$ is the set of functions of the form $\min_{\mathbf{q}_i} \|\mathbf{x} - \mathbf{q}_i\|^2$ parameterized by a set of points $\{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$ where each $\mathbf{q}_i \in \mathbb{R}^d$.

## 3 Wasserstein Coresets

### 3.1 Measure Coresets

Instead of using discrete language, we define a *measure coreset* as a (not necessarily discrete) measure that yields a good approximation to the data distribution under a given family of functions. Given an input measure $\mu$ and a family of functions $\mathcal{F}$, we define a cost in analogy to the discrete case (1) as

$$\text{cost}(\mathcal{X}, f, \mu) = \int_{\mathcal{X}} f(x)\, \mathrm{d}\mu(x), \tag{2}$$

where here $\mathcal{X}$ is no longer a discrete set. With this in mind, we define measure coresets:

**Definition 2** (Measure Coreset)**.** *We call $\nu$ a measure coreset for $\mu$ if $\nu$ is absolutely continuous with respect to $\mu$ and for all $f \in \mathcal{F}$*

$$|\text{cost}(\mathcal{X}, f, \mu) - \text{cost}(\mathcal{X}, f, \nu)| \leq \varepsilon \cdot \text{cost}(\mathcal{X}, f, \mu). \tag{3}$$

Note that such a $\nu$ always exists, since $\nu = \mu$ satisfies the inequality.

**Proposition 1.** *Definition 2 generalizes Definition 1. The latter can be recovered from the former by setting $\mu = \sum_{\mathbf{x} \in X} \mu_X(\mathbf{x}) \delta_{\mathbf{x}}$.*

*Proof.* We verify

$$\text{cost}\left(\mathcal{X}, f, \sum_{\mathbf{x} \in X} \mu_X(\mathbf{x}) \delta_{\mathbf{x}}\right) = \sum_{\mathbf{x} \in X} \mu_X(\mathbf{x}) f(\mathbf{x}).$$

The constraint $\nu \ll \mu$ restricts the support of $\nu$ to a subset $C \subseteq X$. Taking $\mu_C = \nu$ completes the proof. □

For measure coresets to be practical, we restrict to $\nu$ of the form $1/n \sum_{i=1}^n \delta_{\mathbf{x}_i}$. We will henceforth use the absolute error form of (3):

$$|\text{cost}(\mathcal{X}, f, \mu) - \text{cost}(\mathcal{X}, f, \nu)| \leq \varepsilon.$$

Coreset constructions typically use the relative error. However, our construction below yields an absolute error $\varepsilon$-coreset. While not strictly equivalent, absolute error constructions provide stronger guarantees when the cost of the full measure is at least 1, and a weaker guarantee otherwise. To be consistent with prior work, we report relative error in our empirical results. From here on we will use the term $\varepsilon$-coreset to mean *absolute error* $\varepsilon$-measure coreset.

### 3.2 Connection to Optimal Transport

To unravel the connection to optimal transport, we start from the dual problem for the $W_p$ cost:

$$W_1(\mu, \nu) = \sup_{f \in C_b(\mathcal{X})} \int_{\mathcal{X}} f(x) \, \mathrm{d}\mu + \int_{\mathcal{X}} f^c(x) \, \mathrm{d}\nu \tag{4}$$

where $C_b(\mathcal{X})$ is the space of all bounded continuous functions on $\mathcal{X}$, and $f^c(y) = \inf_{x \in \mathcal{X}} \|x - y\|_p^p - f(x)$ is the $c$-transform of $f$.

It is not hard to show in the $p = 1$ case that the constraint $f(x) + f^c(y) \leq |x - y|$ restricts the set of admissible functions to the set of 1-Lipschitz functions in $\mathcal{X}$ and the $c$-transform of $f$ is $-f$ (Santambrogio, 2015; Villani, 2009). We can rewrite (4) as

$$W_1(\mu, \nu) = \left| \sup_{f \in \mathrm{Lip}_1(\mathcal{X})} \int_{\mathcal{X}} f \, \mathrm{d}(\mu - \nu) \right|. \tag{5}$$

From definition (2), we can rewrite the coreset condition as

$$\left| \int_{\mathcal{X}} f(x) \, \mathrm{d}\mu(x) - \int_{\mathcal{X}} f(x) \, \mathrm{d}\nu(x) \right| \leq \varepsilon$$

for all functions $f \in \mathcal{F}$. If $\mathcal{F} \subseteq \mathrm{Lip}_1(\mathcal{X})$, by (5) we obtain a strict upper bound on the coreset error via the Wasserstein cost:

$$\left| \int_{\mathcal{X}} f(x) \, \mathrm{d}\mu(x) - \int_{\mathcal{X}} f(x) \, \mathrm{d}\nu(x) \right| \leq W_1(\mu, \nu)$$

Hence, we have

**Proposition 2** (Wasserstein Coresets)**.** *When $\mathcal{F} \subseteq \mathrm{Lip}_1(\mathcal{X})$, a sufficient condition for $\nu$ to be an $\varepsilon$-coreset for $\mu$ and $\mathcal{F}$ is $W_1(\mu, \nu) \leq \varepsilon$.*

Thus, a strategy for constructing an $n$-point coreset for a measure $\mu$ and a given family of functions $\mathcal{F} \subseteq \mathrm{Lip}_1(\mathcal{X})$ is to solve for $\{\mathbf{x}_i\}_{i=1}^n$ in

$$\operatorname*{arg\,min}_{\mathbf{x}_1, \ldots, \mathbf{x}_n} W_1 \left( \mu, \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{x}_i} \right). \tag{6}$$

Relatively few algorithms have been proposed to solve (6). Instead, we replace $W_1$ in (6) with the 2-Wasserstein cost $W_2$, justified by the inequality $W_1(\mu, \nu) \leq W_2(\mu, \nu)$.

### 3.3 Coreset Construction

We use a recent algorithm for constructing the Wasserstein barycenter of a set of input measures proposed by Claici et al. (2018). We start from the dual problem of $W_2^2 \left(\mu, {}^1\!/_n \sum_{i=1}^n \delta_{\mathbf{x}_i}\right)$. The objective is a function of the points $\{\mathbf{x}_i\}_{i=1}^n$ and the dual potentials $w_i \in \mathbb{R}$:

$$F\left[\{\mathbf{x}_i\}_{i=1}^n, \mathbf{w}; \mu\right] = \frac{1}{n} \sum_{i=1}^n w_i + \sum_{i=1}^n \int_{V_{\mathbf{w}}^i} \left[\|\mathbf{x} - \mathbf{x}_i\|^2 - w_i\right] \mathrm{d}\mu \tag{7}$$

where $V_{\mathbf{w}}^i = \{\mathbf{x} \big| \|\mathbf{x} - \mathbf{x}_i\|^2 - w_i \leq \|\mathbf{x} - \mathbf{x}_j\|^2 - w_j, \forall j\}$ is the power cell of point $\mathbf{x}_i$. The minimizer of (7) with respect to the weight vector $\mathbf{w}$ yields the transport cost $W_2^2 \left(\mu, \frac{1}{n} \sum_{i=1}^n \delta_{\mathbf{x}_i}\right)$ for fixed point positions $\mathbf{x}_i$ (Mérigot, 2011).

Problem (7) is concave in the weights $\mathbf{w}$ and strictly concave modulo constant shifts. However, it is highly non-convex with respect to the point positions. Nonetheless, taking derivatives of $F$ with respect to the weights $w_i$ and points $\mathbf{x}_i$ reveals a simple optimization strategy based on alternating a gradient ascent of a concave problem with a fixed point iteration for the point positions.

Explicitly, we compute

$$\frac{\partial F}{\partial w_i} = \frac{1}{n} - \int_{V_{\mathbf{w}}^i} \mathrm{d}\mu \qquad \frac{\partial F}{\partial \mathbf{x}_i} = \mathbf{x}_i \cdot \int_{V_{\mathbf{w}}^i} \mathrm{d}\mu - \int_{V_{\mathbf{w}}^i} \mathbf{x} \, \mathrm{d}\mu. \tag{8}$$

The complete algorithm is given as Algorithm 1. We use a backtracking search for the step size selection in line 4. As opposed to Claici et al. (2018), we can compute gradients exactly for empirical measures, and thus we do not typically have to rely on an accelerated gradient method.

The crux of the algorithm is in computing the derivatives in (8). We detail how to compute the derivatives for specific cases in section 4.2.

**Algorithm 1** Construct a coreset for $\mu$ with $n$ points.

$\{\mathbf{x}_i\}_{i=1}^n \leftarrow$ random draw of $n$ samples from $\mu$
**for** $t = 1, 2, \ldots, T$ **do**
  **while** $\|\nabla_{\mathbf{w}} F\| > \epsilon$ **do**
    $\alpha \leftarrow \textsc{StepSize}(F, \{\mathbf{x}_i\}_{i=1}^n, \mathbf{w})$
    $\mathbf{w} \leftarrow \mathbf{w} + \alpha \nabla_{\mathbf{w}} F$
  **end while**
  **for** $i = 1, 2, \ldots, n$ **do**
    $\mathbf{x}_i \leftarrow \int_{V_{\mathbf{w}}^i} \mathbf{x} \, \mathrm{d}\mu / \int_{V_{\mathbf{w}}^i} \mathrm{d}\mu$
  **end for**
**end for**

# 4 Analysis

In what follows, we give bounds on the size and construction time of our coreset. These bounds incorporate recent results in approximations of measures by discrete distributions under the Wasserstein metric. In several cases, we match the best known results for deterministic discrete coresets, and empirical results show that our coreset frequently outperforms even randomized constructions.

## 4.1 Coreset Size

We use the following theorem:

**Theorem** (Metric convergence, (Kloeckner, 2012; Brancolini et al., 2009; Weed & Bach, 2017)). *Suppose $\mu$ is a compactly supported measure in $\mathbb{R}^d$ and $\nu_n^*$ is a uniform measure supported on $n$ points that minimizes $W_2^2(\nu_n, \mu)$. Then $W_2^2(\nu_n^*, \mu) \sim \Theta(n^{-1/d})$. Moreover, if $\mu$ is supported on a lower dimensional subspace of $\mathbb{R}^d$ (that is, the support of $\mu$ has Hausdorff dimension $s < d$), then $W_1(\nu_n^*, \mu) \sim \Theta(n^{-1/s})$.*

Our coresets are based on a $W_2$ approximation. Let $\{\mathbf{x}_i\}_{i=1}^n$ minimize $W_2\left(\frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}, \mu\right)$. By Theorem 4.1, and the inequality between $W_p$ metrics, we have the following:

$$\left|\mathrm{cost}(\mathcal{X}, f, \mu) - \mathrm{cost}\left(\mathcal{X}, f, \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}\right)\right| \leq W_1\left(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}\right) \leq W_2\left(\mu, \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}\right).$$

If we choose $n$ large enough such that $W_2\left(\frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}, \mu\right) \leq \varepsilon$, then it also holds that

$$\left|\mathrm{cost}(\mathcal{X}, f, \mu) - \mathrm{cost}\left(\mathcal{X}, f, \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}\right)\right| \leq \varepsilon$$

and thus $\frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$ is an $\varepsilon$-coreset for $\mu$.

Since $W_1\left(\frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}, \mu\right) \sim \Theta(n^{-1/s})$, if we have a globally optimal solution for (6), then the resulting coreset has size $\Theta(\varepsilon^{-s})$. While we cannot guarantee this bound in practice since the approach of Claici et al. (2018) does not guarantee global optimality, empirically we observe that this bound holds and in fact is an overestimate on the size of the coreset.

The size of the coreset is predictably independent of the size of the initial data set as we can generalize to arbitrary measures. Furthermore, the size is independent of additional variables in the underlying problem, e.g. the number of medians in $k$-median clustering.

This improves over the best known deterministic coreset size for $k$-means and $k$-median of $O(k\varepsilon^{-d}\log n)$ (Har-Peled & Mazumdar, 2004), however we must be careful as our coreset bounds are given in absolute error. For $k$-means and $k$-median we are typically in the regime where the full data set has large cost, but if that does not hold, the coresets are no longer comparable. Better randomized construction algorithms exist for both $k$-means/$k$-median and SVM with sizes that do not have such a strong dependence on dimension. Empirically, our coresets are competitive even with randomized construction algorithms.

The previous bounds hold if we can solve (6) to optimality. $F$ is non-convex with respect to the $\mathbf{x}_i$, and problem (7) generalizes $k$-means when $\mu$ is an empirical measure. Hence, solving to optimality is NP-hard. The following weak guarantee, however, follows by definition of $F$:

**Proposition 3** (Weak Guarantee on Coreset Size). *If for fixed $\{\mathbf{x}_i\}_{i=1}^n$ and optimal $\mathbf{w}^*$ it holds that $F[\{\mathbf{x}_i\}_{i=1}^n, \mathbf{w}^*; \mu] \leq \varepsilon$, then $\frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_i}$ is an absolute error $\varepsilon$-coreset for $\mu$.*

### 4.2 Construction Time

Construction time depends strongly on the characteristics of the measure we are approximating. The large majority of the time is spent evaluating the expectations in (8). Since we run the gradient ascent until $\|\nabla_{\mathbf{w}} F\|_2 \leq \epsilon$, and perform $T$ fixed point iterations, the construction requires $O(T/\epsilon)$ calls to an oracle that computes the integrals in (8) for each power cell region $V_w^i$. Since line 8 implements a medoid update step similar to the $k$-means algorithm, we cannot rely on convergence criteria for the outer loop of Algorithm 1. Instead we set a number of maximum iterations $T$. In practice, we observe convergence to a fixed point typically within 20 iterations.

If $\mu = \frac{1}{m}\sum_{i=1}^m \alpha_i \delta_{\mathbf{y}_i}$ is a discrete distribution, then the two expectations can be computed in closed form in time $O(m \log n + dn \log n)$ using a nearest neighbor data structure. To our knowledge, the only other scenario where we can compute (8) in closed form is when $\mu$ is piecewise uniform on finitely many closed and connected regions $X_j$ such that $X = \bigcup_j X_j$. The computation reduces to a convex body volume computation. In this case, the time is dominated by the construction time of the powercell $O(n \log n + n^{\lceil d/2 \rceil})$ (Aurenhammer, 1987).

## 5 Experiments

There are several machine learning problems with Lipschitz cost functions. We test on several machine learning algorithms: $k$-median/$k$-means, SVM binary classification, and linear regression.

### 5.1 Support Vector Machine Training

The hinge loss for a support vector machine is given by

$$l(X, \mathbf{y}; \mathbf{w}) = \sum_{i=1}^m \max(0, 1 - y_i \mathbf{w} \cdot \mathbf{x}_i) \tag{9}$$

where $\mathbf{x}_i, y_i$ are the data points and labels, and $\mathbf{w} \in \mathbb{R}^d$ is a separating hyperplane. The goal is to find a hyperplane $\mathbf{w}$ that linearly separates positive ($y_i = 1$) from negative ($y_i = -1$) examples.

To translate this problem into probabilistic language, define $\mu_+ = \frac{1}{m_+}\sum_{i=1}^m \mathbb{1}_{y_i=1}\delta_{\mathbf{x}_i}$ and $\mu_- = \frac{1}{m_-}\sum_{i=1}^m \mathbb{1}_{y_i=1}\delta_{\mathbf{x}_i}$ where $m_+$ and $m_-$ are appropriate normalization constants. We can parameterize the loss by the hyperplane $\mathbf{w}$, and define a measure theoretic version of (9):

$$l(X; \mathbf{w}) = \int_{\mathcal{X}} \max(0, 1 - \mathbf{w} \cdot \mathbf{x}) \mathrm{d}\mu_+ + \int_{\mathcal{X}} \max(0, 1 + \mathbf{w} \cdot \mathbf{x}) \mathrm{d}\mu_- \tag{10}$$

where we have split the cost over the positive and negative measures. To move from a parameterization over vectors $\mathbf{w}$ to a parameterization over functions, define the function families $\mathcal{F}_+, \mathcal{F}_-$ where

$$\mathcal{F}_\pm = \left\{ f(\mathbf{x}) = \max(0, 1 \mp \mathbf{w} \cdot \mathbf{x}) \;\middle|\; \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_2 \leq \tau \right\}$$

The functions in $\mathcal{F}_\pm$ are Lipschitz with constant $\tau$. We thus incur an error of $\tau \cdot \varepsilon$ which has to be accounted for in coreset construction. Since the function class is Lipschitz we are guaranteed the coreset properties in section 4.1, and can construct a coreset by running Algorithm 1 for $\mu_+$ and $\mu_-$.

To verify our algorithm, we test on the credit card data set of Yeh & Lien (2009) consisting of 24 integer or categorical features. We compare with uniform samples from the empirical distribution, and the Core Vector Machine approach described in (Tsang et al., 2005). Our coreset significantly outperforms both approaches, especially in the regime where the coreset size is small (Figure 1a).

### 5.2 Linear Regression

The least squares cost for a linear estimator given an origin centered data set $X$ is

$$\mathrm{cost}(X; \mathbf{w}) = \|X\mathbf{w}\|_2^2 = \sum_{i=1}^m (\mathbf{w} \cdot \mathbf{x}_i)^2. \tag{11}$$

The least squares error is not Lipschitz. Instead, we use the $l_1$ form of 11. Applying a similar argument to that for $k$-medoids, we can define a cost for measures $\mu$

$$\text{cost}(\mathcal{X}, \mathbf{w}, \mu) = \int_{\mathcal{X}} |\mathbf{w} \cdot \mathbf{x}| \mathrm{d}\mu.$$

The family function $\mathcal{F}$ is here $\{f(\mathbf{x}) = |\mathbf{w} \cdot \mathbf{x}_i| \mid \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w}\|_2 \leq \tau\}$. Since linear regression is efficient even for large data sets in high dimensions, we compare just with the SVD decomposition and a uniform sample of the data (Figure 3b). Coresets for linear regression do exist (Boutsidis et al., 2013), but they essentially compute a low-rank SVD approximation and hence would closely track the performance of SVD.

### 5.3 $k$-Means and $k$-Median

The cost for $k$-medoids for a given set of $k$ centers $Q = \{\mathbf{q}_1, \ldots, \mathbf{q}_k\}$ is given by

$$\text{cost}(X, Q) = \sum_{\mathbf{x} \in X} \min_{i=1,\ldots,k} \|\mathbf{x} - \mathbf{q}_i\|_p^p \tag{12}$$

where $p = 1$ for $k$-median and $p = 2$ for $k$-means. We convert this to measure theoretic language by replacing the discrete set $X$ with a measure $\mu = \frac{1}{n} \sum_{i=1}^m \delta_{\mathbf{x}_i}$:

$$\text{cost}(\mathcal{X}, Q, \mu) = \int_{\mathcal{X}} \min_{i=1,\ldots,k} \|\mathbf{x} - \mathbf{q}_i\|_p^p \mathrm{d}\mu.$$

To translate into the measure coreset language of (3), we define the function families

$$\mathcal{F}_p = \left\{ f(\mathbf{x}) = \min_{\mathbf{q}_i \in Q} \|\mathbf{x} - \mathbf{q}_i\|_p^p \;\middle|\; Q \subset \mathcal{X}, |Q| = k \right\}.$$

The cost for $k$-median is 1-Lipschitz, while the one for $k$-means is not Lipschitz. Recall from (7), however, that we are solving for a minimizer with respect to the $W_2$ transport cost. By setting $w_i = 0$ and taking $\mu = {}^1\!/_m \sum_{j=1}^m \delta_{\mathbf{y}_j}$ a discrete distribution in (7) we recover

$$F\left[\{\mathbf{x}_i\}_{i=1}^n; \mu\right] = \sum_{j=1}^m \min_{i=1,\ldots,n} \|\mathbf{y}_j - \mathbf{x}_i\|^2$$

which is the $k$-means cost for centers $\mathbf{x}_1, \ldots, \mathbf{x}_n$. As $F$ generalizes the $k$-means problem, we expect our approximation to yield an $\varepsilon$-coreset even though the cost is not Lipschitz.

We test our algorithm on two data sets and report results for both $k$-means and $k$-median. The first, shown in Figures 2(a) and 3(a) is a synthetic data set drawn from $k = 10$ normal distributions in $2D$ using $n = 400$ draws from each distribution. The second, shown in Figures 2(b) and 3(b), is the Pendigit data set of 11000 digits represented using $d = 26$ features (Alimoglu & Alpaydin, 1997).

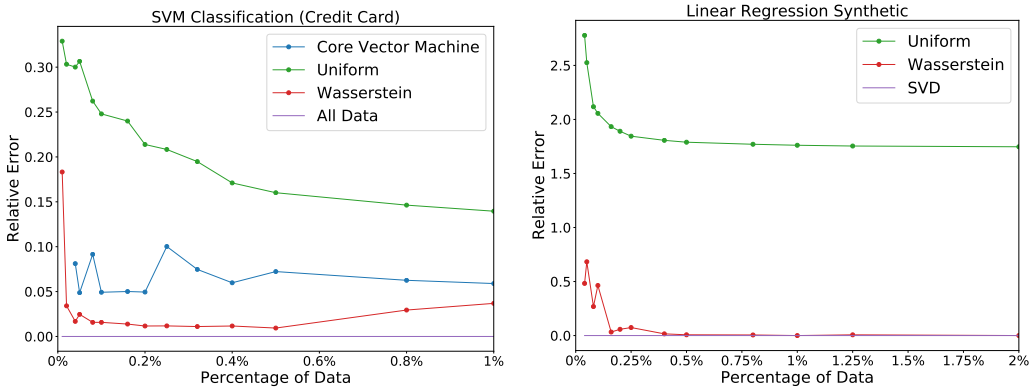
We compare against a uniform sample, a sensitivity approach that uses an $(\alpha, \beta)$-bicriteria and $k$-means++ seeding to bound the sensitivity (Feldman & Langberg, 2011), and a deterministic algorithm designed for the streaming setting (Barger & Feldman, 2016). Our algorithm outperforms competing methods for both $k$-means and $k$-median clustering (Figures 2 and 3).

## 6 Discussion

We have introduced *measure coresets*, an extension of discrete coresets to probability measures. Our construction reveals a surprising connection to optimal transport theory which leads to a simple construction algorithm that applies to any problem with Lipschitz cost.

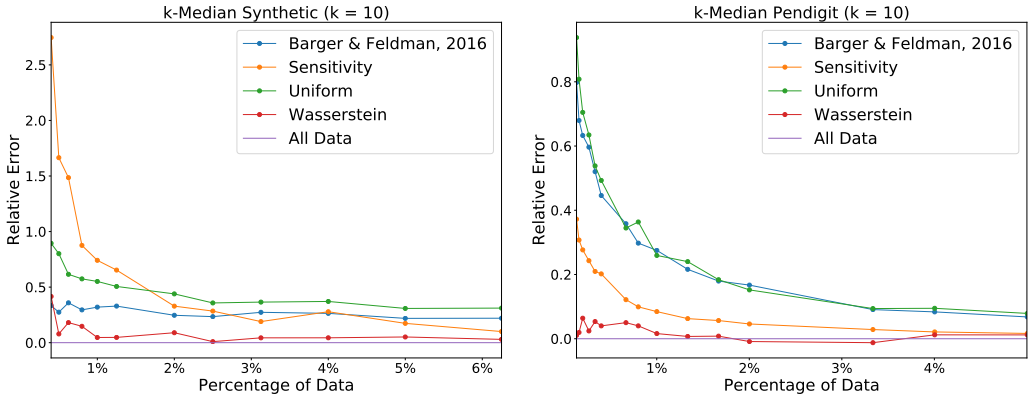
The behavior of our algorithm reveals several avenues for future research. We highlight specifically the discrepancy between the theoretical guarantees of our algorithm and its empirical performance. This gap is large and its existence indicates that stronger bounds can be proved about our coreset construction. Specifically, we conjecture that $\text{Lip}(\mathcal{X})$ is a much too conservative function class as we solve for a $W_2$ approximation, instead of $W_1$.

More generally, while we have used the Wasserstein metric to find a coreset, our *measure coresets* are defined independently of the construction presented here. We expect that more efficient algorithms exist for constructing such coresets using either other metrics or different approximation building blocks (Gaussian instead of discrete for example).



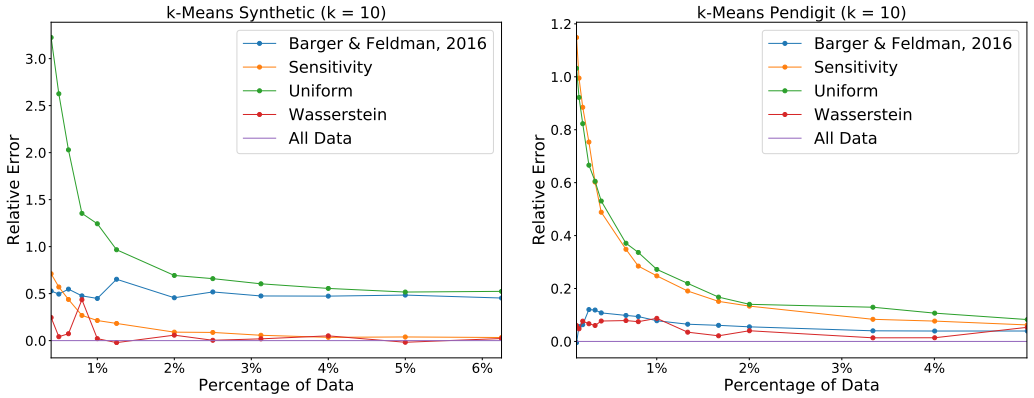

(a) (b)

Figure 1: Classification and regression. (a) Evaluation of the SVM coreset against a uniform sample, and the Core Vector Machine approach of Tsang et al. (2005). Note that our coreset performs well even with only a fraction of the data. (b) Comparison with SVD and a uniform sample on a synthetic data set. Our coreset accurately recovers the principal components with only a fraction of the data.




(a) (b)

Figure 2: Evaluation of $k$-median coreset against the construction from Barger & Feldman (2016), a sensitivity approach, and uniform sampling. (a) Synthetic data set drawn from 10 multivariate normal distributions with $k = 10$. (b) Pendigit data set with $k = 10$ clusters.




(a) (b)

Figure 3: Evaluation of $k$-means coreset against the construction of Barger & Feldman (2016), a sensitivity approach, and uniform sampling. (a, b) As in Figure 2. Despite the cost not being Lipschitz, our performance is comparable to that of Barger & Feldman (2016).

## References


Agarwal, Pankaj K, Har-Peled, Sariel, and Varadarajan, Kasturi R. Geometric approximation via coresets. *Combinatorial and computational geometry*, 52:1–30, 2005.

Alimoglu, Fevzi and Alpaydin, Ethem. Combining multiple representations and classifiers for pen-based handwritten digit recognition. In *4th International Conference Document Analysis and Recognition (ICDAR '97), 2-Volume Set, August 18-20, 1997, Ulm, Germany, Proceedings*, pp. 637–640, 1997. doi: 10.1109/ICDAR.1997.620583. URL https://doi.org/10.1109/ICDAR.1997.620583.

Arjovsky, Martín, Chintala, Soumith, and Bottou, Léon. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 214–223, 2017. URL http://proceedings.mlr.press/v70/arjovsky17a.html.

Arthur, David and Vassilvitskii, Sergei. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007*, pp. 1027–1035. Society for Industrial and Applied Mathematics, 2007.

Aurenhammer, Franz. Power diagrams: properties, algorithms and applications. *SIAM Journal on Computing*, 16(1):78–96, 1987.

Bachem, Olivier, Lucic, Mario, and Lattanzi, Silvio. One-shot coresets: The case of k-clustering. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pp. 784–792, 2018. URL http://proceedings.mlr.press/v84/bachem18a.html.

Barger, Artem and Feldman, Dan. *k*-means for streaming and distributed big sparse data. In *Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, Florida, USA, May 5-7, 2016*, pp. 342–350, 2016. doi: 10.1137/1.9781611974348.39. URL https://doi.org/10.1137/1.9781611974348.39.

Baykal, Cenk, Liebenwein, Lucas, Gilitschenski, Igor, Feldman, Dan, and Rus, Daniela. Data-dependent coresets for compressing neural networks with applications to generalization bounds. *CoRR*, abs/1804.05345, 2018. URL http://arxiv.org/abs/1804.05345.

Boutsidis, Christos, Drineas, Petros, and Magdon-Ismail, Malik. Near-optimal coresets for least-squares regression. *IEEE Trans. Information Theory*, 59(10):6880–6892, 2013. doi: 10.1109/TIT.2013.2272457. URL https://doi.org/10.1109/TIT.2013.2272457.

Brancolini, Alessio, Buttazzo, Giuseppe, Santambrogio, Filippo, and Stepanov, Eugene. Long-term planning versus short-term planning in the asymptotical location problem. *ESAIM: Control, Optimisation and Calculus of Variations*, 15(3):509–524, 2009.

Campbell, Trevor and Broderick, Tamara. Bayesian coreset construction via greedy iterative geodesic ascent. *CoRR*, abs/1802.01737, 2018. URL http://arxiv.org/abs/1802.01737.

Carrière, Mathieu, Cuturi, Marco, and Oudot, Steve. Sliced wasserstein kernel for persistence diagrams. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 664–673, 2017. URL http://proceedings.mlr.press/v70/carriere17a.html.

Claici, Sebastian, Chien, Edward, and Solomon, Justin. Stochastic Wasserstein barycenters. *Proceedings of the 35th International Conference on Machine Learning, ICML 2018 (to appear)*, abs/1802.05757, 2018. URL http://arxiv.org/abs/1802.05757.

Cuturi, Marco. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013*, pp. 2292–2300, 2013. URL http://papers.nips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optimal-transport.

De Goes, Fernando, Breeden, Katherine, Ostromoukhov, Victor, and Desbrun, Mathieu. Blue noise through optimal transport. *ACM Transactions on Graphics (TOG)*, 31(6):171, 2012.

Feldman, Dan and Langberg, Michael. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pp. 569–578, 2011. doi: 10.1145/1993636.1993712. URL http://doi.acm.org/10.1145/1993636.1993712.

Feldman, Dan, Schmidt, Melanie, and Sohler, Christian. Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2013*, pp. 1434–1453. SIAM, 2013.

Genevay, Aude, Peyré, Gabriel, and Cuturi, Marco. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617, 2018.

Har-Peled, Sariel and Mazumdar, Soham. On coresets for k-means and k-median clustering. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing, STOC 2004*, pp. 291–300. ACM, 2004.

Huggins, Jonathan H., Campbell, Trevor, and Broderick, Tamara. Coresets for scalable bayesian logistic regression. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4080–4088, 2016. URL http://papers.nips.cc/paper/6486-coresets-for-scalable-bayesian-logistic-regression.

Kloeckner, Benoît. Approximation by finitely supported measures. *ESAIM Control Optim. Calc. Var.*, 18(2):343–359, 2012. ISSN 1292-8119.

Langberg, Michael and Schulman, Leonard J. Universal epsilon-approximators for integrals. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2010, Austin, Texas, USA, January 17-19, 2010*, pp. 598–607, 2010. doi: 10.1137/1.9781611973075.50. URL https://doi.org/10.1137/1.9781611973075.50.

Lévy, Bruno. A Numerical Algorithm for L2 Semi-Discrete Optimal Transport in 3D. *ESAIM Math. Model. Numer. Anal.*, 49(6):1693–1715, November 2015. ISSN 0764-583X, 1290-3841. doi: 10.1051/m2an/2015055.

Lévy, Bruno and Schwindt, Erica. Notions of optimal transport theory and how to implement them on a computer. *arXiv:1710.02634*, 2017.

Mérigot, Quentin. A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pp. 1583–1592. Wiley Online Library, 2011.

Montavon, Grégoire, Müller, Klaus-Robert, and Cuturi, Marco. Wasserstein training of restricted boltzmann machines. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3711–3719, 2016. URL http://papers.nips.cc/paper/6248-wasserstein-training-of-restricted-boltzmann-machines.

Munteanu, Alexander and Schwiegelshohn, Chris. Coresets-methods and history: A theoreticians design pattern for approximation and streaming algorithms. *Künstliche Intelligenz (KI)*, 32(1): 37–53, 2018.

Peyré, Gabriel and Cuturi, Marco. *Computational Optimal Transport*. Submitted, 2018.

Santambrogio, Filippo. *Optimal Transport for Applied Mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer International Publishing, Cham, 2015. ISBN 978-3-319-20827-5 978-3-319-20828-2. doi: 10.1007/978-3-319-20828-2.

Schmitz, Morgan A., Heitz, Matthieu, Bonneel, Nicolas, Mboula, Fred Maurice Ngolè, Coeurjolly, David, Cuturi, Marco, Peyré, Gabriel, and Starck, Jean-Luc. Wasserstein dictionary learning: Optimal transport-based unsupervised non-linear dictionary learning. *CoRR*, abs/1708.01955, 2017. URL http://arxiv.org/abs/1708.01955.

Solomon, Justin. *Optimal Transport on Discrete Domains*. AMS Short Course on Discrete Differential Geometry, 2018.

Srivastava, Sanvesh, Cevher, Volkan, Dinh, Quoc, and Dunson, David. WASP: Scalable Bayes via barycenters of subset posteriors. In Lebanon, Guy and Vishwanathan, S. V. N. (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 912–920, San Diego, California, USA, 09–12 May 2015. PMLR. URL http://proceedings.mlr.press/v38/srivastava15.html.

Staib, Matthew, Claici, Sebastian, Solomon, Justin M, and Jegelka, Stefanie. Parallel streaming Wasserstein barycenters. In *Advances in Neural Information Processing Systems, NIPS 2017*, pp. 2644–2655, 2017.

Tsang, Ivor W., Kwok, James T., and Cheung, Pak-Ming. Core vector machines: Fast SVM training on very large data sets. *Journal of Machine Learning Research*, 6:363–392, 2005. URL http://www.jmlr.org/papers/v6/tsang05a.html.

Villani, Cédric. *Optimal Transport: Old and New*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009. ISBN 978-3-540-71049-3. OCLC: ocn244421231.

Weed, Jonathan and Bach, Francis. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *CoRR*, abs/1707.00087, 2017. URL http://arxiv.org/abs/1707.00087.

Yeh, I-Cheng and Lien, Che-hui. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Syst. Appl.*, 36(2):2473–2480, 2009. doi: 10.1016/j.eswa.2007.12.020. URL https://doi.org/10.1016/j.eswa.2007.12.020.