Unconditional Coresets for Regularized Loss Minimization

Ryan R. Curtin
RelationalAI
ryan@ratml.org

Sungjin Im
University of California Merced
sim3@ucmerced.edu

Benjamin Moseley
Carnegie Mellon University
moseleyb@andrew.cmu.edu

Kirk Pruhs
University of Pittsburgh
kirk@cs.pitt.edu

Alireza Samadian
University of Pittsburgh
samadian@cs.pitt.edu

Abstract

We design and mathematically analyze sampling-based algorithms for regularized loss minimization problems that are implementable in popular computational models for large data, in which the access to the data is restricted in some way. Our main result is that if the regularizer's effect does not become negligible as the norm of the hypothesis scales, and as the data scales, then a uniform sample of modest size is with high probability a coreset. In the case that the loss function is either logistic regression or soft-margin support vector machines, and the regularizer is one of the common recommended choices, this result implies that a uniform sample of size $O(d\sqrt{n})$ is with high probability a coreset of n points in $\Re^d$. We contrast this upper bound with two lower bounds. The first lower bound shows that our analysis of uniform sampling is tight; that is, a smaller uniform sample will likely not be a core set. The second lower bound shows that in some sense uniform sampling is close to optimal, as significantly smaller core sets do not generally exist.

1 Introduction

We consider the design and mathematical analysis of sampling-based algorithms for regularized loss minimization (RLM) problems on large data sets (Shalev-Shwartz and Ben-David, 2014). The input consists of a collection $X = \{x_1, x_2, \ldots, x_n\}$ of points in $\Re^d$, and a collection $Y = \{y_1, y_2, \ldots, y_n\}$ of associated labels from $\{-1, 1\}$. Intuitively the goal is to find a hypothesis $\beta \in \Re^d$ that is the best "linear" explanation for the labels. More formally, the objective is to minimize a function $F(\beta)$ that is a linear combination of a nonnegative nondecreasing loss function ℓ that measures the goodness of the hypothesis, and a nonnegative regularization function r that measures the complexity of the hypothesis. So:

$$F(\beta) = \sum_{i=1}^{n} \ell(-y_i \beta \cdot x_i) + \lambda\, r(R\beta) \tag{1}$$

Notable examples include regularized logistic regression, where the loss function is $\ell(z) = \log(1 + \exp(z))$, and regularized soft margin support vector machines (SVM), where the loss function is $\ell(z) = \max(0, 1+z)$. Common regularizers are the 1-norm, the 2-norm, and the 2-norm squared (Buhlmann and van de Geer, 2011). The parameter $\lambda \in \Re$ is ideally set to balance the risks of over-fitting and under-fitting. We will assume that λ is proportional to n^κ for some $0 < \kappa < 1$, capturing the range of most commonly suggested regularizers. In particular, is commonly recommended to set λ to be proportional to $\Theta(\sqrt{n})$ (Shalev-Shwartz and Ben-David, 2014; Negahban et al., 2009). For this choice of λ, if there was a true underlying distribution from which the data was drawn in an i.i.d. manner, then there is a guarantee that the computed β will likely have vanishing relative error with respect to the ground truth (Shalev-Shwartz and Ben-David, 2014, Corollary 13.9) (Negahban et al., 2009, Corollary 3). The parameter R is the maximum 2-norm of any point in X. Note that the regularizer must scale with R if it is to avoid having a vanishing effect as the point set X scales.[1]

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

[1] To see this note that if we multiplied each coordinate of each point x_i by a factor of c, the optimal hypothesis β would decrease by a factor of c, thus decreasing the value of all of the standard regularizers.

We are particularly interested in settings where the data set is too large to fit within the main memory of one computer, and thus the algorithm's access to the data set is restricted in some way. Popular computation models that arise from such settings include:

Streaming Model: This model derives from settings where the data is generated in real-time, or stored on a memory technology (such as a disk or tape) where a sequential scan is way more efficient than random accesses. In this model, the data can only be accessed by a single (or a small number of) sequential passes (Muthukrishnan, 2005).

Massively Parallel Computation (MPC) Model: This model derives from settings where the data is distributed over multiple computers. In this model only a few rounds of communication with sublinear sized messages are allowed (Im et al., 2017; Karloff et al., 2010).

Relational Model: This model derives from settings where the data is stored in a database in a collection of tables. In this model the data must be accessed via relational operators that do not explicitly join tables (Khamis et al., 2016).

Thus we additionally seek algorithms that can be reasonably implemented in these popular restricted access models.

One popular method to deal with large data sets is to extract a manageably small (potentially weighted) sample from the data set, and then directly solve (a weighted version of) the RLM problem on the (weighted) sample[2]. The aspiration here is that the optimal solution on the sample will be a good approximation to the optimal solution on the original data set. To achieve this aspiration, the probability that a particular point is sampled (and the weight that it is given) may need to be carefully computed as some points may be more important than other points. But if this sampling probability distribution is too complicated, it may not be efficiently implementable in common restricted access models.

A particularly strong condition on the sample that is sufficient for achieving this aspiration is that the sample is a *coreset*; intuitively, a sample is a coreset if *for all possible hypotheses* β, the objective value of β on the sample is very close to the objective value of β on the whole data set.

There has been work on constructing coresets for special cases of the RLM problem. In particular, sublinear coresets exists for *unregularized* logistic regression (i.e $\lambda = 0$) by making assumptions on the input. The exact assumption is technical, but intuitively the coresets are small when there is no hypothesis that is a good explanation of the labels. The work of Tolochinsky and Feldman (2018) gave coresets for regularized soft-margin SVM assuming the 2-norm of the optimal β is small.

Unfortunately, both of these works do not apply to general input instances. Moreover, they require knowledge that is not easy to compute about the input to know if the input has the needed properties. One may wonder if small coresets exist for general data sets. The work of Munteanu et al. (2018) shows that there is no coreset of size $\Omega(\frac{n}{\log n})$ for unregularized logistic regression.

This lower bound is discouraging, suggesting that small coresets are not possible for arbitrary inputs even for the special case of the logistic regression problem. However, the lower bound is for unregularized logistic regression. In practice, regularization is almost always used, as emphasized in the following quotes. From Chapter 5. Basic Practice of Burkov (2019): "Regularization is the most widely used approach to prevent overfitting." Quoting Maya Gupta, head of the Glassbox Machine Learning team at Google from her online course on machine learning, "The key ingredient to making machine learning work great... is regularization" (Gupta).

Thus, the natural research question is if small coresets exist for RLM problems in general, and for regularized logistic regression and regularized SVM – and further, if they can be efficiently computed within the common restricted access models.

Our Results: Our main result is that if the regularizer's effect does not become negligible as the norm of the hypothesis scales then a uniform sample of size $\Theta(n^{1-\kappa}\Delta)$ points is with high probability a coreset. Here, Δ is the VC-dimension of the loss function. Thus, coresets exists for general input instances for the RLM problem, showing regularization allows us to break through lower bounds shown in prior work! Formally this scaling condition says that if $\ell(-\|\beta\|) = 0$ then $r(\beta)$ must be a constant fraction of $\ell(\|\beta\|_2)$. We show that this scaling condition holds when the loss function is either logistic regression or SVM, and the regularizer is the 1-norm, the 2-norm, or the 2-norm squared. So for example, in the recommended case that $\kappa = 1/2$, the scaling condition ensures that a uniform sample of $\tilde{\Theta}(d\sqrt{n})$ points is with high probability a coreset when the regularizer is one of the standard ones, and the loss function is either logistic regression and SVM, as they have VC-dimension $O(d)$. Note also

[2]A more general approach is to summarize the data set in some more sophisticated way than as a weighted sample, but such approaches are beyond the scope of this paper.

that uniform sampling can be reasonably implemented in all of the popular restricted access models. So this yields a reasonable algorithm for all of the restricted access models under the assumption that a data set of size $\tilde{\Theta}(d\sqrt{n})$ can be stored, and reasonably solved, in the main memory of one computer.

We complement our upper bound with two lower bounds on the size of coresets. Our lower bounds assume the 2-norm squared as the regularizer, since intuitively this is the standard regularizer for which it should be easiest to attain small coresets. We first show that our analysis is asymptotically tight for uniform sampling. That is, we show that for both logistic regression and SVM, a uniform sample of size $O(n^{1-\kappa-\epsilon})$ may not result in a coreset. We then show for both logistic regression and SVM there are instances in which every core set is of size $\Omega(n^{(1-\kappa)/5-\epsilon})$. So more sophisticated sampling methods must still have core sets whose size is in the same ballpark as is needed for uniform sampling. One might arguably summarize our results as saying that the simplest possible sampling method is nearly optimal for obtaining a coreset.

We experimentally evaluate the practical utility of uniform sampling for logistic regression using several real-world datasets from the UCI machine learning dataset repository (Dua and Graff, 2017). We observe that our theory is empirically validated as uniform samples yield good empirical approximation, and orders-of-magnitude speedup over learning on the full dataset.

Related Work on Coresets: The most closely related prior work is probably Munteanu et al. (2018), who considered coresets for *unregularized* logistic regression; i.e, the regularization parameter $\lambda = 0$. Munteanu et al. (2018) showed that are data sets for which there do not exist coresets of sublinear size, and then introduced a parameter μ of the instances that intuitively is small when there is no hypothesis that is a good explanation of the labels, and showed that a coreset of size roughly linear in μ can be obtain by sampling each point with a uniform probability plus a probability proportional to its ℓ_2^2 leverage scores (which can be computed from a singular value decomposition of the points). This result yields an algorithm, for the promise problem in which μ is known a priori to be small (but it is not clear how to reasonably compute μ), that is reasonably implementable in the MPC model, and with two passes over the data in the streaming model. It seems unlikely that this algorithm is implementable in the relational model due to the complex nature of required sampling probabilities. Contemporaneously with our research, Tolochinsky and Feldman (2018) obtained results similar in flavor to those of Munteanu et al. (2018). Tolochinsky and Feldman (2018) also show that small coresets exist for certain types of RLM instances; in this case, those in which the norm of the optimal hypothesis is small. So for normalized logistic regression Tolochinsky and Feldman (2018) shows that when the 2-norm of the optimal β is bound by μ, coresets of size $\tilde{O}(\mu^2 n^{1-\kappa})$ can be obtained by sampling a point with probability proportional to its norm divided by its ordinal position in the sorted order of norms. So again this yields an algorithm for the promise problem in which μ is known a priori to be small (and again it is not clear how to reasonably compute μ). Due to the complex nature of the probabilities it is not clear that this algorithm is reasonably implementable in any of the restricted access models that we consider. So from our perspective there are three key differences between the results of Munteanu et al. (2018) and Tolochinsky and Feldman (2018) and our positive result: *(1)* our result applies to all data sets *(2)* we use uniform sampling, and thus *(3)* our sampling algorithm is implementable in all of the restricted access models that we consider.

Surveys of the use of coresets in algorithmic design can be found in Munteanu and Schwiegelshohn (2018) and in Har-Peled (2011, Chapter 23). The knowledge that sampling with probability at least proportional to sensitivity yields a coreset has been used for at least a decade as it is used by Dasgupta et al. (2009). Coresets were used for partitioned clustering problems, such as k-means (Har-Peled and Mazumdar, 2004; Meyerson et al., 2004; Bachem et al., 2018). Coresets for hard margin SVM are known (Har-Peled et al., 2007). These coresets have an approximation guarantee on the quality of the margin to the hyperplane. Unfortunately, these ideas not not applicable to soft-margin SVM.

Coresets have been used the Minimum Enclosing Ball (MEB) problem (Har-Peled, 2011). Coresets for MEB are the basis for the Core Vector Machine approach to unregularized kernelized SVM (Tsang et al., 2005). Several strong coresets for computing balls are known (Bădoiu and Clarkson, 2008; Badoiu and Clarkson, 2003). We note that while there is a reduction from kernelized SVM to MEB, the reduction is not approximation preserving, and thus the existence of coresets for MEB does not imply the existence of coresets for SVM.

Coresets have also been used for submodular optimization Mirrokni and Zadimoghaddam (2015), clustering Badoiu et al. (2002), Baysian Logistic Regression Huggins et al. (2016) and in the design of streaming algorithms (e.g. O'Callaghan et al. (2002)), as well as distributed algorithms (e.g. Malkomes et al. (2015)).

2 Preliminaries

We define $\ell_i(\beta) = \ell(-y_i\beta \cdot x_i)$ as the contribution of point i to the loss function. We define $f_i(\beta) = \ell(-y_i\beta \cdot x_i) + \lambda r(R\beta)/n$ as the contribution of point i to the objective $F(\beta)$. The sensitivity of point i is then $s_i = \sup_\beta \ f_i(\beta)/F(\beta)$, and the total sensitivity is $S = \sum_{i=1}^n s_i$. For $\epsilon > 0$, an ϵ-coreset (C, U) consists of a subcollection C of $[1, n]$, and associated nonnegative weights $U = \{u_i \mid i \in C\}$, such that

$$\forall \beta \qquad H(\beta) \coloneqq \frac{\left|\sum_{i=1}^n f_i(\beta) - \sum_{i\in C} u_i f_i(\beta)\right|}{\sum_{i=1}^n f_i(\beta)} \leq \epsilon \quad (2)$$

Conceptually one should think of u_i as a multiplicity, that is that x_i is representing u_i points from the original data set. So one would expect that $\sum_{i\in C} u_i = n$, although this is not strictly required. But it is easy to observe that $\sum_{i\in C} u_i$ must be close to n.

Observation 1. *Assume that $\ell(0) \neq 0$, as is the case for logistic regression and SVM. If (C, U) is an ϵ-coreset then $(1-\epsilon)n \leq \sum_{i\in C} u_i \leq (1+\epsilon)n$.*

Proof. Applying the definition of coreset in the case that β is the hypothesis with all 0 components, it must be the case that $\left|\sum_{i=1}^n \ell(0) - \sum_{i\in C} u_i\ell(0)\right| \leq \epsilon \sum_{i=1}^n \ell(0)$, or equivalently $\left|n - \sum_{i\in C} u_i\right| \leq \epsilon n$. □

Note that in the special case that each u_i is equal to a common value u, as will be the case for uniform sampling, setting each $u_i = 1$ and scaling λ down by a factor of u, would result in the same optimal hypothesis β.

A collection X of data points is *shatterable* by a loss function ℓ if for every possible set of assignments of labels, there is a hypothesis β and a threshold t, such that for the positively labeled points $x_i \in X$ it is the case the $\ell(\beta \cdot x_i) \geq t$, and for the negatively labeled points x_i it is the case that $\ell(\beta \cdot x_i) < t$. The VC-dimension of a loss function is then the maximum cardinality of a shatterable set. It is well known that if the loci of points $x \in \Re^d$ where $\ell(\beta \cdot x) = t$ is a hyperplane then the VC-dimension is at most $d+1$ (Vapnik, 1998). It is obvious that this property holds if the loss function is SVM, and Munteanu and Schwiegelshohn (2018) show that it holds if the loss function is logistic regression. The regularizer does not affect the VC-dimension of a RLM problem.

A loss function ℓ and a regularizer r satisfy the (σ, τ)-scaling condition if $\ell(-\sigma) > 0$, and if $\|\beta\|_2 \geq \sigma$ then $r(\beta) \geq \tau\ \ell(\|\beta\|_2)$. Intuitively this condition ensures that the objective value of any correctly classified point that is near the separating hyperplane must be bounded away from zero, that is either the loss function or the regularlizer must bounded away from zero.

Theorem 2 (Feldman and Langberg (2011); Braverman et al. (2016))**.** *Let $(n, X, Y, \ell, r, \lambda, R, \kappa)$ be an instance of the RLM problem where the loss function has VC-dimension at most Δ. Let s'_i be an upper bound on the sensitivity s_i, let $S' = \sum_{i=1}^n s'_i$. Let $\epsilon, \delta \in (0, 1)$ be arbitrary. Let C be a random sample of at least $\frac{10S'}{\epsilon^2}(\Delta \log S' + \log(\frac{1}{\delta}))$ points sampled in an i.i.d fashion, where the probability that point $i \in [1, n]$ is selected each time is s'_i/S'. Let the associated weight u_i for each point $x_i \in C$ be $\frac{S'}{s'_i|C|}$. Then C and $U = \{u_i \mid x_i \in C\}$ is an ϵ-coreset with probability at least $(1-\delta)$.*

3 Upper Bound for Uniform Sampling

In this section, we show that uniform sampling can be used to construct a coreset for regularized loss minimization.

Theorem 3. *Let $(n, X, Y, \ell, r, \lambda, R, \kappa)$ be an instance of the RLM problem where ℓ and r satisfy the (σ, τ)-scaling condition and the loss function has VC-dimension at most Δ. Let $S' = \frac{n}{\tau\lambda} + \frac{\ell(\sigma)}{\ell(-\sigma)} + 1$. A uniform sample of $q = \frac{10S'}{\epsilon^2}(\Delta \log S' + \log(\frac{1}{\delta}))$ points, each with an associated weight of $u = n/q$, is an ϵ-coreset with probability at least $1 - \delta$.*

Proof. With an aim towards applying Theorem 2 we start by upper bounding the sensitivity of an arbitrary point. To this end consider an arbitrary $i \in [1, n]$ and an arbitrary hypothesis β. First consider the case that $R\|\beta\|_2 \geq \sigma$. In this case:

$$\begin{aligned}
\frac{f_i(\beta)}{F(\beta)} &= \frac{\ell(-y_i\beta \cdot x_i) + \frac{\lambda}{n} r(R\beta)}{\sum_j \ell(-y_j\beta \cdot x_j)) + \lambda\, r(R\beta)} \\
&\leq \frac{\ell(|\beta \cdot x_i|) + \frac{\lambda}{n} r(R\beta)}{\sum_j \ell(-y_j\beta \cdot x_j) + \lambda\, r(R\beta)} \\
&\qquad \text{[As the loss function is nondecreasing]} \\
&\leq \frac{\ell(|\beta \cdot x_i|) + \frac{\lambda}{n} r(R\beta)}{\lambda\, r(R\beta)} \\
&\qquad \text{[As the loss function is nonnegative]} \\
&\leq \frac{\ell(|\beta \cdot \beta| \frac{R}{\|\beta\|_2}) + \frac{\lambda}{n} r(R\beta)}{\lambda\, r(R\beta)} \\
&\qquad \text{[As maximum is when } x_i = \beta \frac{R}{\|\beta\|_2}\text{]} \\
&\leq \frac{\ell(R\ \|\beta\|_2)}{\lambda\, r(R\beta)} + \frac{1}{n} \\
&\leq \frac{\ell(R\ \|\beta\|_2)}{\lambda\, \tau\, \ell(R\,\|\beta\|_2)} + \frac{1}{n} \\
&\qquad \text{[By } (\sigma, \tau) \text{ scaling assumption} \\
&\qquad \text{and assumption } R\,\|\beta\|_2 \geq \sigma\text{]}
\end{aligned}$$

$$\leq \frac{1}{\tau\lambda} + \frac{1}{n}$$

Next consider the case that $R\|\beta\|_2 < \sigma$. In this case:

$$\begin{aligned}
\frac{f_i(\beta)}{F(\beta)} &= \frac{\ell(-y_i\beta\cdot x_i) + \frac{\lambda}{n}r(R\beta)}{\sum_j \ell(-y_j\beta\cdot x_j) + \lambda\, r(R\beta)} \\
&\leq \frac{\ell(|\beta\cdot x_i|) + \frac{\lambda}{n}r(R\beta)}{\sum_j \ell(-|\beta\cdot x_j|) + \lambda r(R\beta)} \\
&\quad \text{[As the loss function is nondecreasing]} \\
&\leq \frac{\ell(|\beta\cdot\beta|\frac{R}{\|\beta\|_2}) + \frac{\lambda}{n}r(R\beta)}{\sum_j \ell(-|\beta\cdot\beta|\frac{R}{\|\beta\|_2}) + \lambda\, r(R\beta)} \\
&\quad \text{[As maximum is when } x_i = \beta\frac{R}{\|\beta\|_2}\text{]} \\
&\leq \frac{\ell(R\|\beta\|_2) + \frac{\lambda}{n}r(R\beta)}{\sum_j \ell(-R\|\beta\|_2) + \lambda\, r(R\beta)} \\
&\leq \frac{\ell(R\ \|\beta\|_2)}{\sum_j \ell(-R\ \|\beta\|_2)} + \frac{1}{n} \\
&\quad \text{[As } a,b,c,d \geq 0 \text{ implies } \frac{a+b}{c+d} \leq \frac{a}{c} + \frac{b}{d}\text{]} \\
&\leq \frac{\ell(\sigma)}{\sum_j \ell(-\sigma)} + \frac{1}{n} \\
&\quad \text{[By assumption } R\|\beta\|_2 < \sigma\text{]} \\
&\leq \frac{\ell(\sigma)}{n\,\ell(-\sigma)} + \frac{1}{n}
\end{aligned}$$

Thus the sensitivity of every point is at most $\frac{1}{\tau\lambda} + \frac{\ell(\sigma)}{n\,\ell(-\sigma)} + \frac{1}{n}$, and the total sensitivity S is at most $\frac{n}{\tau\lambda} + \frac{\ell(\sigma)}{\ell(-\sigma)} + 1$. The claim the follows by Theorem 2. □

Corollary 4. *Let $(n, X, Y, \ell, r, \lambda, R, \kappa)$ be an instance of the RLM problem where the loss function ℓ is logistic regression or SVM, and the regularizer r is one of the 1-norm, 2-norm, or 2-norm squared. Let $S' = \frac{12n}{\lambda} + 6 = 12n^{1-\kappa} + 6$. A uniform sample of $q = \frac{10S'}{\epsilon^2}((d+1)\log S' + \log(\frac{1}{\delta}))$ points, each with an associate weight of $u = \frac{n}{q}$, is an ϵ-coreset with probability at least $1-\delta$.*

Proof. Since the VC-dimension of logistic regression and SVM is at most $d+1$, it is enough to show that the scaling condition holds in each case. First consider logistic regression. Let $\sigma = 1$. Then we have $l(-1) = \log(1+\exp(-1)) \neq 0$. In the case that $r(\beta) = \|\beta\|_2$ it is sufficient to take $\tau = \frac{1}{2}$ as $\ell(z) = \log(1+\exp(z)) \leq 2z$ when $z \geq 1$. Similarly its sufficient to take $\tau = \frac{1}{2}$ when the regularizer is the 2-norm squared, as $\ell(z) = \log(1+\exp(z)) \leq 2z^2$ when $z \geq 1$. As $\|\beta\|_1 \geq \|\beta\|_2$ it is also sufficient to take $\tau = \frac{1}{2}$ when the regularizer is the 1-norm. Therefore, total sensitivity is bounded by $\frac{2n}{\lambda} + 6$ in all of these cases.

Now consider SVM. Let $\sigma = 1/2$. Then $l(-1/2) = 1/2 \neq 0$. In the case that $r(\beta) = \|\beta\|_2$ it is sufficient to take $\tau = \frac{1}{3}$ as $\ell(z) = 1 + z \leq 3z$ when $z \geq \frac{1}{2}$; $\tau = \frac{1}{3}$ will be also sufficient when the regularizer is the 1-norm since $\|\beta\|_1 \geq \|\beta\|_2$.

Furthermore, if $\|\beta\|_2 \geq 1$, then $\|\beta\|_2^2 \geq 4\|\beta\|_2$; therefore, in the case that $r(\beta) = \|\beta\|_2^2$, it is sufficient to take $\tau = \frac{1}{12}$. Therefore, total sensitivity is bounded by $\frac{12n}{\lambda} + 4$. □

The implementation of uniform sampling, and the computation of R, in the streaming and MPC models is trivial. Uniform sampling and the computation of R in the relational model can be implemented without joins because both can be expressed using *functional aggregate queries*, which can then be efficiently computed without joins (Khamis et al., 2016).

We note that in several other papers (e.g. Munteanu and Schwiegelshohn (2018)) coreset constructions can be applied recursively to obtain very small coresets. We cannot apply the previous theorem recursively because after sampling and re-weighting the regularizer stays the same; however, the number of points is less and the weight of loss function for each point is scaled. To see that it is not possible to re-sample the new instance, it is enough to divide the new error function by the weight of each sample and get an unweighted instance that its regularizer has a small coefficient; now, having a coreset for the weighted sample is similar to having a coreset for this unweighted sample with small regularizer. Of course, it can also be seen that the theorem cannot be applied recursively because it would contradict out lower bound in the next section as well.

4 Uniform Sampling Lower Bound

In this section we show in Theorem 5 that our analysis of uniform sampling is tight up to poly-logarithmic factors.

Theorem 5. *Assume that the loss function is either logistic regression or SVM, and the regularizer is the 2-norm squared. Let $\epsilon, \gamma \in (0,1)$ be arbitrary. For all sufficiently large n, there exists an instance I_n of n points such that with probability at least $1 - 1/n^{\gamma/2}$ it will be the case that for a uniform sample C of $c = n^{1-\gamma}/\lambda = n^{1-\kappa-\gamma}$ points, there is no weighting U that will result in an ϵ-coreset.*

Proof. The instance I_n consists of points located on the real line, so the dimension $d = 1$. A collection A of $n - (\lambda n^{\gamma/2})$ points is located at $+1$, and the remaining $\lambda n^{\gamma/2}$ points are located at -1; call this collection of points B. All points are labeled $+1$. Note $R = 1$.

Let C be the random sample of c points, and U an arbitrary weighting of the points in C. Note that U may depend on the instantiation of C. Our goal is to show that with high probability, (C, U) is not an ϵ-coreset. Our proof strategy is to first show that because almost all of the points are in A, it is likely that C contains only points from A. Then we want to show that, conditioned on $C \subseteq A$, that C can not be a coreset for any possible weighting. We accomplish this by showing that $\lim_{n\to\infty} H(\beta) = 1$ when $\beta = n^{\gamma/4}$. The details are relatively straightforward and can be found in the appendix. □

5 General Lower Bound on Coreset Size

This section is devoted to proving the following theorem for logistic regression (the proof for SVM is given in the appendix).

Theorem 6. *Assume that the loss function is either logistic regression or SVM, and the regularizer is the 2-norm squared. Let $\epsilon, \gamma \in (0,1)$ be arbitrary. For all sufficiently large n, there exists an instance I_n of n points such that I_n does not have an ϵ-coreset of size $O(n^{(1-\kappa)/5-\gamma})$.*

The lower bound instance I_n consists of a collection of n positively-labeled points in $\Re^3$ uniformly spaced around a circle of radius 1 centered at $(0,0,1)$ in the plane $z = 1$. Note that $R = \sqrt{2}$. However for convenience, we will project I_n down into a collection X of points in the plane $z = 0$. So the resulting instance, which we call the circle instance, consists of n points uniformly spread around the unit circle in $\Re^2$. So for a hypothesis $\beta = (\beta_x, \beta_y, \beta_z)$, $F(\beta)$ is now $\sum_{x_i \in X} \ell(-y_i((\beta_x, \beta_y) \cdot x_i + \beta_z)) + 2\lambda \|\beta\|_2^2$. So β_z can be thought of as an offset or bias term, that allows hypotheses in $\Re^2$ that do not pass through the origin.

Fix a constant $c > 0$ and a subset C of X that has size $k = c\frac{n^{1/5-\gamma}}{\lambda^{1/5}} = cn^{(1-\kappa)/5-\gamma}$ as a candidate coreset. Let U be an arbitrary collection of associated weights. Toward finding a hypothesis that violates equation (2), define a *chunk* A to be a collection of $\frac{n}{4k}$ points in the middle of $\frac{n}{2k}$ consecutive points on the circle that are all not in C. So no point in the chunk A is in C, and no point in the next $\frac{n}{8k}$ points in either direction around the circle are in C. Its easy to observe that, by the pigeon principle, a chunk A must exist. Now let $\beta_A = (\beta_x, \beta_y, \beta_z)$ be the hypothesis where $(\beta_x, \beta_y) \cdot x_i + \beta_z = 0$ for the two points $x_i \in X \setminus A$ that are adjacent to the chunk A, that predicts A incorrectly (and thus that predicts the points $X \setminus A$ correctly), and where $\|\beta_A\|_2 = \sqrt{\frac{n^{1-\gamma}}{k\lambda}}$. To establish Theorem 6 we want to show that equation (2) is not satisfied for the hypothesis β_A. By Observation 1 it is sufficient to show that the limit as $n \to \infty$ of:

$$\frac{\left|\sum_{x_i \in X} \ell_i(\beta_A) - \sum_{x_i \in C} u_i \ell_i(\beta_A)\right| - 2\epsilon\lambda \|\beta_A\|_2^2}{\sum_{x_i \in X} \ell_i(\beta_A) + \lambda \|\beta_A\|_2^2} = \frac{\left|1 - \frac{\sum_{x_i \in C} u_i \ell_i(\beta_A)}{\sum_{x_i \in X} \ell_i(\beta_A)}\right| - \frac{2\epsilon\lambda\|\beta_A\|_2^2}{\sum_{x_i \in X} \ell_i(\beta_A)}}{1 + \frac{\lambda\|\beta_A\|_2^2}{\sum_{x_i \in X} \ell_i(\beta_A)}}$$

is 1. To accomplish this it is sufficient to show that the limits of the ratios in the second expression approach 0, which we do in the next two lemmas.

Lemma 7. $\lim_{n\to\infty} \frac{\lambda\|\beta_A\|_2^2}{\sum_{x_i \in X} \ell_i(\beta_A)} = 0.$

Proof. As the $\frac{n}{4k}$ points in A have been incorrectly classified by β_A, we know that $\ell_i(\beta_A) \geq \log 2$ for $x_i \in A$. Thus:

$$\begin{aligned}\lim_{n\to\infty} \frac{\lambda \|\beta_A\|_2^2}{\sum_{x_i \in X} \ell_i(\beta_A)} &\leq \lim_{n\to\infty} \frac{\lambda \frac{n^{1-\gamma}}{k\lambda}}{\frac{n}{4k} \log 2} \\ &= \lim_{n\to\infty} \frac{4}{n^\gamma \log 2} \\ &= 0.\end{aligned}$$

□

Lemma 8. $\lim_{n\to\infty} \frac{\sum_{x_i \in C} u_i \ell_i(\beta_A)}{\sum_{x_i \in X} \ell_i(\beta_A)} = 0.$

Proof. Let d_i be the distance between x_i and the line that passes through the first and last points in the chunk A. Let θ_i be the angle formed by the the ray from the origin through x_i and the ray from the origin to them middle point in A. Let $\theta = \max_{i \in A} \theta_i = \frac{2\pi}{n}\frac{n}{8k} = \frac{\pi}{4k}$. We then make two algebraic observations (the proof of the first can be found in the appendix, and the second is more or less obvious).

Observation 9. *For all $x_i \in X$, $d_i \|\beta_A\|_2 / 2 \leq |(\beta_x, \beta_y) \cdot x_i + \beta_z| \leq d_i \|\beta_A\|_2$.*

Observation 10. *For all $x_i \in X$, $d_i = |\cos(\theta_i) - \cos(\theta)|$.*

We then have:

$$\lim_{n\to\infty} \frac{\sum_{x_i \in C} u_i \ell_i(\beta_A)}{\sum_{x_i \in X} \ell_i(\beta_A)}$$

dataset	n	d	coreset error $\overline{H}(\beta)$ $\lvert C\rvert = \sqrt{n}$	$\lvert C\rvert = 10\sqrt{n}$	$\lvert C\rvert = 20\sqrt{n}$
`connect4`	67557	126	0.35 ± 0.01	0.04 ± 0.00	0.02 ± 0.00
`grid_stability`	10000	12	0.52 ± 0.00	0.02 ± 0.00	0.01 ± 0.00
`miniboone`	130064	50	0.78 ± 0.01	0.39 ± 0.00	0.22 ± 0.00
`mnist`	70000	784	0.53 ± 0.02	0.19 ± 0.02	0.14 ± 0.00
`pokerhand`	1000000	85	1.01 ± 0.00	0.00 ± 0.00	0.00 ± 0.00

Table 1: Dataset information and relative approximation of logistic regression objective with coresets C of different sizes. Three trials are used. Coreset error of 0 indicates a very good approximation.

$$= \lim_{n\to\infty} \frac{\sum_{x_i\in C} u_i \log\left(1+\exp(-((\beta_x,\beta_y)\cdot x_i+\beta_z))\right)}{\sum_{x_i\in X}\ell_i(\beta_A)}$$

$$\le \lim_{n\to\infty} \frac{\sum_{x_i\in C} u_i \log\left(1+\exp(-\frac{d_i\|\beta_A\|_2}{2})\right)}{\sum_{x_i\in A}\ell_i(\beta_A)}$$

[By Observation 9]

$$\le \lim_{n\to\infty} \frac{\sum_{x_i\in C} u_i \log\left(1+\exp(-\frac{\|\beta_A\|_2}{2}(\cos\theta-\cos\theta_i))\right)}{\sum_{x_i\in A}\ell_i(\beta_A)}$$

[By Observation 10]

$$\le \lim_{n\to\infty} \frac{\sum_{x_i\in C} u_i \exp(-\frac{\|\beta_A\|_2}{2}(\cos\theta-\cos\theta_i))}{\sum_{x_i\in A}\ell_i(\beta_A)}$$

[Since $\log(1+x)\le x$]

$$\le \lim_{n\to\infty} \frac{\sum_{x_i\in C} u_i \exp(-\frac{\|\beta_A\|_2}{2}(\cos\frac{\pi}{4k}-\cos\frac{\pi}{2k}))}{\sum_{x_i\in A}\ell_i(\beta_A)}$$

[Since maximizer is when $\theta_i = \frac{\pi}{2k}$]

$$\le 0$$

[As $\cos\frac{\pi}{4k}-\cos\frac{\pi}{2k} = \frac{3\pi^2}{32k^2} - O(\frac{1}{k^4})$]

More details can be found in Appendix B.1. □

6 Experiments

We next experimentally evaluate the practical utility of our uniform sampling scheme for logistic regression. Using 5 datasets from the UCI machine learning dataset repository (Dua and Graff, 2017), we uniformly generate samples of different sizes and train a logistic regression model.

These datasets are collected from synthetic and real-world data sources, and so represent a reasonable collection of diverse datasets. Table 1 gives details on the number of points (n) and the number of dimensions (d) for each dataset.

Logistic regression models are trained using the mlpack C++ machine learning library (Curtin et al., 2018).

We first validate the general efficacy of uniform random sampling by running three trials with three different coreset sizes: $\sqrt{n}$, $10\sqrt{n}$, and $20\sqrt{n}$. We plot the relative difference in loss measures (0 means a perfect approximation). Specifically, the approximation given in the table, $H(\beta)$, is given as

$$H(\beta) = \frac{\left|\sum_{i=1}^n f_i(\beta) - \sum_{i\in C} u_i f_i(\beta)\right|}{\sum_{i=1}^n f_i(\beta)}. \tag{3}$$

We report the mean of $H(\beta)$ over three trials: $\overline{H}(\beta)$.

Next, we graph the approximation error of uniform random sampling as the sample size is swept from 50 points up to the dataset size. We use $\lambda = 0.1$, and report the mean approximation $H(\beta)$ (according to Eqn. 2) in Figure 1. When training the models, the L-BFGS optimizer is used until convergence (Liu and Nocedal, 1989). However, although this works for our experiments, note that in general it is not feasible to use L-BFGS like this, specifically when datasets are very large, or when we are in restricted computation access models, as we have considered in this paper. This is because a single L-BFGS step requires computation of the gradient of $f_i(\beta)$ for *every* $i \in [n]$.

In extremely-large-data or streaming settings, a typical strategy for training a logistic regression model is to use mini-batch SGD (Ruder, 2016), where the model's parameters are iteratively updated using a gradient computation on a small batch of random points. However, SGD-like optimizers can converge very slowly in practice and have a number of parameters to configure (learning rate, number of epochs, batch size, and so forth). But because our theory allows us to choose a sufficiently small sample, we can use a full-batch optimizer like L-BFGS and this often converges to a much

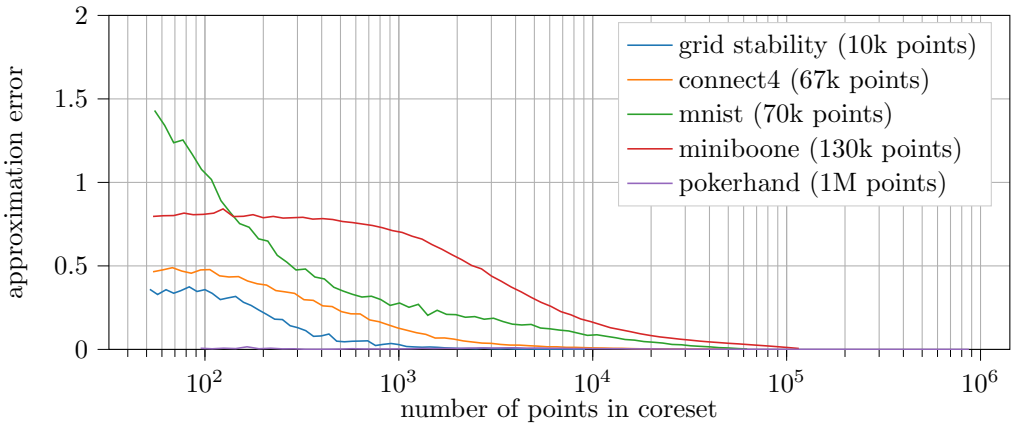
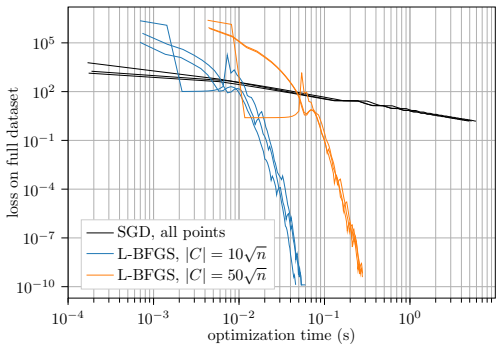


Figure 1: sample approximation error vs. sample size for different datasets.

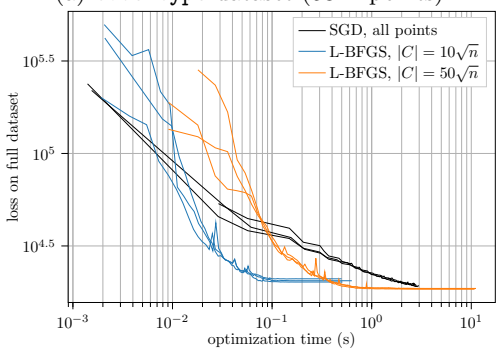
better solution orders of magnitude more quickly.

To demonstrate this, we train a logistic regression model on a sample using L-BFGS and on the full dataset using SGD for 20 epochs. At each step of the optimization, we record the wall-clock time and compute the loss on the full training set (the loss computation time is not included in the wall-clock time). Figure 2 shows the results for three trials of each strategy on two moderately-sized datasets. It is clear from these results that a full-batch gradient descent technique can provide a good approximation of the full-dataset model with orders-of-magnitude speedup; in fact, L-BFGS is often able to recover a much better model than even 20 epochs of SGD!

Overall, we can see that our theory is empirically validated: uniform sampling provides samples that give good empirical approximation, and the use of these samples can result in orders-of-magnitude speedup for learning models. Thus, our theory shows and our experiments justify that uniform sampling to obtain coresets is a compelling and practical approach for restricted access computation models.



(a) `covertype` dataset (581k points).



(b) `mnist` dataset (70k points).

Figure 2: Learning curves; log-log axes. Three trials of each strategy are shown. Note the orders-of-magnitude faster convergence for L-BFGS on samples.

7 Conclusion

This paper considered constructing coresets for regularized loss minimization problems. We gave an algorithm that constructs a coreset. The algorithm is essentially the best possible, ensuring the coreset has size that nearly matches a lower bound shown in the paper. The algorithm is simple and easy to implement in most large data models.

8 Acknowledgments

S. Im was supported in part by NSF grants CCF-1617653 and CCF-1844939. B. Moseley was supported in part by NSF grants CCF-1725543, 1733873, 1845146, a Google Research Award, a Bosch junior faculty chair and an Infor faculty award. K. Pruhs was supported in part by NSF grants CCF-1421508 and CCF-1535755, and an IBM Faculty Award.

References

Olivier Bachem, Mario Lucic, and Andreas Krause. Scalable k -means clustering via lightweight coresets. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2018.

Mihai Badoiu and Kenneth L. Clarkson. Smaller coresets for balls. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms, January 12-14, 2003, Baltimore, Maryland, USA.*, pages 801–802, 2003.

Mihai Badoiu, Sariel Har-Peled, and Piotr Indyk. Approximate clustering via core-sets. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing, May 19-21, 2002, Montréal, Québec, Canada*, pages 250–257, 2002.

Vladimir Braverman, Dan Feldman, and Harry Lang. New frameworks for offline and streaming coreset constructions. *CoRR*, abs/1612.00889, 2016.

Mihai Bădoiu and Kenneth L. Clarkson. Optimal coresets for balls. *Comput. Geom. Theory Appl.*, 40(1): 14–22, May 2008. ISSN 0925-7721.

Peter Buhlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Publishing Company, 2011.

Andriy Burkov. *The Hundred-Page Machine Learning Book*. Andriy Burkov, 2019.

Ryan R. Curtin, Marcus Edel, Mikhail Lozhnikov, Yannis Mentekidis, Sumedh Ghaisas, and Shangtong Zhang. mlpack 3: a fast, flexible machine learning library. *Journal of Open Source Software*, 3:726, 2018. URL `https://doi.org/10.21105/joss.00726`.

Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W. Mahoney. Sampling algorithms and coresets for ℓ_p regression. *SIAM J. Comput.*, 38(5):2060–2078, 2009.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *ACM Symposium on the Theory of Computing*, 2011.

Maya Gupta. Machine learning crash course. URL `https://developers.google.com/machine-learning/crash-course/regularization-for-simplicity/video-lecture`.

S. Har-Peled and S. Mazumdar. On coresets for k-means and k-median clustering. In *ACM-SIAM Symposium on Discrete Algorithms*, 2004.

Sariel Har-Peled. *Geometric Approximation Algorithms*. American Mathematical Society, 2011.

Sariel Har-Peled, Dan Roth, and Dav Zimak. Maximum margin coresets for active and noise tolerant learning. In *IJCAI 2007, Proceedings of the 20th International Joint Conference on Artificial Intelligence, Hyderabad, India, January 6-12, 2007*, pages 836–841, 2007.

Jonathan H. Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable bayesian logistic regression. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4087–4095, 2016. ISBN 978-1-5108-3881-9.

Sungjin Im, Benjamin Moseley, and Xiaorui Sun. Efficient massively parallel methods for dynamic programming. In *ACM Symposium on the Theory of Computing*, 2017.

Howard J. Karloff, Siddharth Suri, and Sergei Vassilvitskii. A model of computation for mapreduce. In *ACM-SIAM Symposium on Discrete Algorithms*, 2010.

Mahmoud Abo Khamis, Hung Q. Ngo, and Atri Rudra. FAQ: questions asked frequently. In *ACM Symposium on Principles of Database Systems*, 2016.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

Gustavo Malkomes, Matt J. Kusner, Wenlin Chen, Kilian Q. Weinberger, and Benjamin Moseley. Fast distributed k-center clustering with outliers on massive data. In *NeuralPS*, 2015.

Adam Meyerson, Liadan O'Callaghan, and Serge A. Plotkin. A k-median algorithm with running time independent of data size. *Machine Learning*, 56(1-3):61–87, 2004.

Vahab S. Mirrokni and Morteza Zadimoghaddam. Randomized composable core-sets for distributed submodular maximization. In *ACM Symposium on the Theory of Computing*, 2015.

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

Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David P. Woodruff. On coresets for logistic regression. In *NeurIPS*, 2018.

S. Muthukrishnan. Data streams: Algorithms and applications. *Foundation and Trends in Theoretical Computer Science*, 1(2):117–236, August 2005.

Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Neural Information Processing Systems*, pages 1348–1356, 2009.

Liadan O'Callaghan, Adam Meyerson, Rajeev Motwani, Nina Mishra, and Sudipto Guha. Streaming-data algorithms for high-quality clustering. In *International Conference on Data Engineering*, pages 685–694, 2002.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Elad Tolochinsky and Dan Feldman. Coresets for monotonic functions with applications to deep learning. *CoRR*, abs/1802.07382, 2018. URL `http://arxiv.org/abs/1802.07382`.

Ivor W. Tsang, James T. Kwok, and Pak-Ming Cheung. Core vector machines: Fast svm training on very large data sets. *Journal of Machine Learning Research*, 6:363–392, 2005.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.