Clustering High-dimensional Data with Ordered Weighted ℓ_1 Regularization

Chandramauli Chakraborty* Indian Statistical Institute, Kolkata, India

Sayan Paul* Indian Statistical Institute, Kolkata, India

Saptarshi Chakraborty Department of Statistics UC Berkeley

Swagatam Das Electronics and Communication Sciences Unit, Indian Statistical Institute

Abstract

Clustering complex high-dimensional data is particularly challenging as the signal-to-noise ratio in such data is significantly lower than their classical counterparts. This is mainly because most of the features describing a data point have little to no information about the natural grouping of the data. Filtering such features is, thus, critical in harnessing meaningful information from such large-scale data. Many recent methods have attempted to find feature importance in a centroid-based clustering setting. Though empirically successful in classical low-dimensional settings, most perform poorly, especially on microarray and single-cell RNA-seq data. This paper extends the merits of weighted center-based clustering through the Ordered Weighted ℓ_1 (OWL) norm for better feature selection. Appealing to the elegant properties of block coordinate-descent and Frank-Wolf algorithms, we are not only able to maintain computational efficiency but also able to outperform the state-of-the-art in high-dimensional settings. The proposal also comes with finite sample theoretical guarantees, including a rate of $\mathcal{O}\left(\sqrt{k \log p/n}\right)$, under model-sparsity, bridging the gap between theory and practice of weighted clustering.

1 Introduction

Clustering is one of the main concepts in unsupervised machine learning, where one has data but no labels. The goal is to partition the data points into subsets so that each group's data points share some typical pattern or characteristics. The data points are usually represented using a vector of some measurement (commonly referred to as *features*). Often in real-world data, the clusters only express themselves in a handful of features in the entire feature space. Finding out this subset of features is critical in understanding the natural grouping of the data.

Out of the plethora of methods used in clustering, perhaps the most widely used ones fall under the category of center-based hard clustering, where each cluster is represented by its centroid. Even after 60 years of its inception, the classical k-means algorithm with Lloyd'd heuristic (Macqueen, 1967; Jain, 2010) is the most common approach for hard center-based clustering mostly due to its fastness and interpretability. Given a dataset, $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, k-means tries to minimize the within-cluster variance of the data points by minimizing the following objective function

$$f_{k-means}(\Theta) = \sum_{i=1}^{n} \min_{1 \le j \le k} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|_2^2 \qquad (1)$$

where $\Theta = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$ is the set of the k centroids and $\|\cdot\|_2$ is the usual ℓ_2 norm. The objective function (1) can be minimized using Lloyd's algorithm (Lloyd, 1982), which uses a two-step alternating minimization procedure. k-measn have been further generalized to model-based clustering from a statistical perspective mainly using a mixture of distributions (McNicholas, 2016; Fraley and Raftery, 1998; McLachlan and Rathnayake, 2014) as well as from a Bayesian viewpoint (Kulis and Jordan, 2011).

Technological advancements have made it simpler to obtain enormous amounts of real-world data that are described using thousands of attributes, thus giving rise to high-dimensional data. For instance, photographs can have billions of pixels, text, and web articles can have thousands of words, and microarray datasets can have thousands of genes' expression levels. The term "curse of dimensionality" (Bellman, 2003) is frequently used to refer to some fundamental issues with high-dimensional data, where $p \gg n$ and the concept of the nearest neighbors fade out effectively makes the standard Euclidean distance (Beyer et al.,

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

*Co-first authors contributed equally

Corresponding author: Swagatam Das, email: swagatam.das@isical.ac.in

1999) extremely ineffective. Additionally, many scholars agree that, particularly for high-dimensional data, meaningful clusters may only be present in select subspaces constructed using a particular subset of the features available (Tsai and Chiu, 2008; Liu and Yu, 2005; Chen et al., 2012). The challenge is exacerbated by the fact that various features may show varying degrees of relevance to the underlying groups. To handle this issue, machine learning algorithms typically use a variety of ways to choose or ignore features. When a significant number of features are not relevant to some clusters, using all the features available can reduce the accuracy of the final clustering solutions and even confound the learning algorithm used for the cluster analysis (and generally for any pattern recognition task) (Chan et al., 2004). It's common to think of feature weighting (Chan et al., 2004; DeSarbo et al., 1984; Li and Yu, 2006) as a generalisation of the widely-used feature selection techniques (Wettschereck et al., 1997; Modha and Spangler, 2003; de Amorim, 2016). Recently, the concepts of feature-weighting and feature-ranking have been successfully used to tackle this curse of dimensionality both for classical and high-dimensional data. For examples, see the works of Witten and Tibshirani (2010); Chakraborty and Das (2020); Chakraborty et al. (2020); Zhang et al. (2020).

From a practical viewpoint, we ask, *(Q_1)"Does equal feature importance (in revealing the cluster structure) imply equal feature weight?"* In particular, we want to know if correlated features give rise to similar feature weights. For all the above methods, the answer is no. Though unimportant features will often get a zero feature weight, especially if one uses some sort of weight penalization as used by Witten and Tibshirani (2010) or Chakraborty and Das (2020), the feature weights will often not be the same for essential features, even if the features are equally important and have identical distributions.

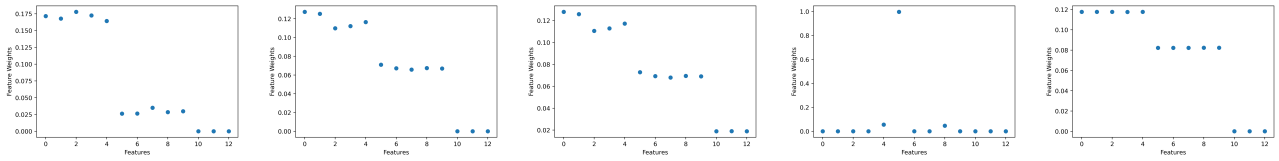
The second question we must ask is that Q_2 *"What are the theoretical advantages of using such feature weighting?"*. In the classical regression literature, it is well known (for a rigorous treatment, see chapter 7 of Wainwright (2019)) that using an ℓ_1 penalization greatly depletes the excess risk from $\mathcal{O}\left(\sqrt{\frac{p}{n}}\right)$ to $\mathcal{O}\left(\sqrt{\frac{\log p}{n}}\right)$ under model sparsity. In the clustering literature, such a guarantee is not yet known.

To solve the first question, inspired by the recent successful application of Ordered Weighted ℓ_1 (OWL) norm in linear regressions (Bogdan et al., 2015; Bao et al., 2020) to address regression problem with correlated covariates. The OWL norm can be viewed as a generalization of the ℓ_1 norm and has intriguing properties to facilitate feature selection and also has applications in multiple testing problems. Our objective function uses the OWL norm as a penalty term to encourage correlated features to give equal importance when the dissimilarity is measured in a weighted distance. The resulting objective function is solved through an alternating combination of Frank-Wolf (Frank and Wolfe, 1956) and coordinate descent for weight and centroid updates, respectively. From a statistical viewpoint, we analyze the finite-sample properties of the proposal. We show that one can indeed give an affirmative answer to Q_2, thus making the clustering guarantees at par as its regression counterpart.

Our main contributions can be summarized as follows:

- After going over some necessary concepts in Section 2, we provide a straightforward sparse clustering framework based on feature weighting called Ordered Weighted ℓ_1-k-means (OWL-k-means) clustering, where an *owl* penalty is explicitly applied to the feature weights, in Section 3.
- We employ an alternative block-coordinate descent and Frank-Wolfe algorithm (Frank and Wolfe, 1956) in section 3 to minimize the proposed objective function. Closed-form updates are produced by the process, which keeps the simplicity of Lloyd's algorithm Lloyd (1982).
- The theoretical finite-sample properties of the (global) solutions the objective is thoroughly analyzed in Section 4 through the aid of tools in learning theory such as Rademacher complexities. We analytically show that the excess risk scales as $\mathcal{O}\left(\sqrt{k \log p/n}\right)$ under model sparsity. Such logarithmic rates in the dimensions have not previously been observed in the clustering literature.
- The efficacy of OWL-k-means is thoroughly demonstrated through simulations and in-depth experiments on microarray gene expression and single-cell RNA sequence data in Sections 5 and 6, followed by concluding remarks in Section 7.

A Motivating Example As a motivating example, we present a case study on synthetic data with three distinct levels of feature importance. We generate a data set with $n = 200$ observation in two clusters, where the first five features are most important, and the next five are relatively less important. The last three are completely unrelated to the cluster structure of the data. On this example we run different feature-weighting-based methods such as Entropy Weighted Power k-means (EWP) (Chakraborty et al., 2020), Lasso Weighted k-means (LWK) (Chakraborty and Das, 2020), Weighted k-means (Huang et al., 2005), Sparse k-means (Witten and Tibshirani, 2010) alongside our proposal. In Fig. 1, we plot the different feature weights obtained by the algorithms. All the methods except sparse k-means can reflect the feature importance of the correlated features; only OWL-k-means can infer that the correlated



(a) EWP (b) LWK (c) Weighted-k-means (d) Sparse-k-means (e) OWL-k-means (ours)

Figure 1: Feature weights obtained by the peer algorithms with three different levels of feature variances. OWL-k-means properly identifies that correlated features have equal feature importance.

features effectively carry the same level of cluster information.

2 Background and Preliminaries

Notations $\mathbb{R}^p_{\geq 0}$ denotes the set of all non-negative real vectors of length p. For any vector $\boldsymbol{w} \in \mathbb{R}^p_{\geq 0}$, $\|\boldsymbol{x}\|^2_{\boldsymbol{w}} = \sum_{\ell=1}^p w_\ell x_\ell^2$, for any vector $\boldsymbol{x} \in \mathbb{R}^p$. For any $m \in \mathbb{N}$, $[m] = \{1, \ldots, m\}$. For any vector, $\boldsymbol{w} \in \mathbb{R}^p$, $\boldsymbol{w}^\beta := (w_1^\beta, \ldots, w_p^\beta)$. For any vector $\boldsymbol{a}$, $(\boldsymbol{a})_\ell$ denotes the ℓ-th coordinate of $\boldsymbol{a}$.

Weighted k-means Huang et al. (2005) developed the Weighted k-means (W-k-means) clustering as a minimization of the following objective function to include feature weighting in the k-means type clustering:

$$f_{W-k-means}(\Theta, \boldsymbol{w}) = \sum_{i=1}^n \min_{1\leq j\leq k} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2_{\boldsymbol{w}^\beta} \qquad (2)$$

where, $\boldsymbol{w} = (w_1, \ldots, w_p) \in \mathbb{R}^p_{\geq 0}$ is the feature weight vector, under the constraint $\sum_{l=1}^p w_l = 1$. The hyper-parameter, $\beta > 1$, is provided by the user. A block-coordinate descent heuristic is employed to locally minimize the objective function (2). The technique maintains Lloyd's k-means version's computational simplicity. Some noteworthy extensions of W-k-means can be found in the works of Jing et al. (2007); Huang et al. (2007); De Amorim and Mirkin (2012); Chakraborty et al. (2020) among many others.

Ordered Weighted ℓ_1 (OWL) norm Bogdan et al. (2015) developed an ordered weighted ℓ_1 norm, a generalization of the ℓ_1 norm, which they have used for sparse regression and variable selection, inspired by ideas in multiple testing of the linear regression estimator. The OWL norm of $\boldsymbol{w}$ is defined as,

$$\Omega_{\boldsymbol{\lambda}}(\boldsymbol{w}) = \sum_{i=1}^p \lambda_i |w|_{(i)}, \qquad (3)$$

where, $0 \leq \lambda_1 \leq \cdots \leq \lambda_p$ and $|w|_{(1)} \leq \cdots \leq |w|_{(p)}$ are the increasing sequence of the absolute value of the $\boldsymbol{w}$ vector. If all the λ_i's are identical, then we get the usual ℓ_1 norm. The OSCAR regularizer proposed by Bondell and Reich (2008) is a special case of OWL. There are many choices of the λ_i's, but one popular choice suggested by Bogdan et al. (2015) is

$$\lambda_{p-i+1} = \Phi^{-1}(1 - q_i), \qquad q_i = i\frac{q}{2p},\ q \in (0, 1) \qquad (4)$$

Here q is a parameter that must be provided by the user. Here, the Φ is the CDF of the standard normal distribution function. The above choice has amicable relations with BH-testing (Benjamini and Hochberg, 1995) as shown in Bogdan et al. (2015).

Frank-Wolf Algorithm Frank-Wolfe algorithm (Frank and Wolfe, 1956) is a celebrated technique for constrained convex optimization. Consider the problem

$$\min_{\boldsymbol{x}\in\mathcal{C}} g(\boldsymbol{x}) \qquad (5)$$

where $\mathcal{C}$ is a convex set. Under some preferable assumptions, the Frank-Wolfe algorithm solves this problem by considering a linear approximation of the objective function. Using an auxiliary variable, the method approaches a minimizer of this linear function. This algorithm guarantees the convergence to the minimizer at a sublinear rate. Some special variants of the Frank-Wolfe can be found in the works of Lacoste-Julien and Jaggi (2015); Kunisch and Walter (2021); Cristofari et al. (2017) among many others.

3 OWL-k-means

Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be the data matrix, where p is the number of features, n is the number of data points and $\boldsymbol{x}_i$ denote the i-th data point (i.e the i-th row of $\boldsymbol{X}$). Let $\boldsymbol{\Theta} \in \mathbb{R}^{k\times p}$ be the matrix of centroids, where $\boldsymbol{\theta}_j$ denote the i-th centroid (i.e. the j-th row of $\boldsymbol{\Theta}$). Here, $k \in \mathbb{N}$ is the number of clusters, which is assumed to be known to the user. The OWL-k-means objective is defined as,

$$f(\boldsymbol{\Theta}, \boldsymbol{w}) = \sum_{i=1}^n \min_{1\leq j\leq k} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2_{\boldsymbol{w}^\beta} + \Omega_{\boldsymbol{\lambda}}(\boldsymbol{w}). \qquad (6)$$

This objective is minimized subject to $w_i \geq 0$, $\boldsymbol{w}^T\mathbf{1} = 1$. Here, $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_p$. The variable $\boldsymbol{w} \in \mathbb{R}^p$ is interpreted as the vector of feature weights. The first term

of the equation (6), $\sum_{i=1}^{n} \min_{1\le j\le k} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2_{\boldsymbol{w}^\beta}$ measures the fit of the cluster centroids in the $\|\cdot\|_{\boldsymbol{w}^\beta}$-norm. For instance say $w_\ell \propto 1$, then the first term reduces to a scalar multiple of $\sum_{i=1}^{n} \min_{1\le j\le k} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|_2^2$, which is the objective function of the classical k-means algorithm, and is equivalent to the within-cluster variance. If we remove the constraint of $w_i \propto 1$, then we are applying weights to each feature, and the centroid is chosen depending on these weighted within-cluster variances. This fit term thus guides the cluster centroids to find a good clustering in terms of the weighted distance that reflects the feature importance. The feature weight $\boldsymbol{w}$ enables us to learn which features are significantly helpful for distinguishing these clusters. Instead of associating a linear weighting framework (i.e., $\|\cdot\|_{\boldsymbol{w}}$ as used by Chakraborty et al. (2020)) to the within-cluster variance, we have incorporated a power of weights to generalize the problem further.

The second term of the equation (6) is the penalty term applied to the feature weights. This penalty term encourages features with similar within-cluster variances i.e., $D_\ell = \sum_{j=1}^{k} \sum_{i\in C_j} (x_{i\ell} - \theta_{jl})^2$ to have equal feature weight and noisy features which have a typical large within-cluster sum of squares to give rise to a zero weight. Here, C_j denotes the j-the cluster. We have used this owl norm to generalize the concept of using an ℓ_1 norm like Witten and Tibshirani (2010); Chakraborty and Das (2020). This norm penalizes features with a higher variance that is a higher D_l value, which is why the weight corresponding to l-th feature is associated with a higher λ term to penalize that feature further.

Optimization The objective function (6) can be reformulated as

$$f(U, \boldsymbol{\Theta}, \boldsymbol{w}) = \sum_{i=1}^{n}\sum_{j=1}^{k} u_{ij}\|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2_{\boldsymbol{w}^\beta} + \sum_{\ell=1}^{p} \lambda_\ell w_{(\ell)} \quad (7)$$

such that,

$$u_{ij} \in [0,1], \text{ for all } i = 1,\dots,n;\ j = 1,\dots,k,$$
$$\sum_{j=1}^{k} u_{ij} = 1, \text{ for all } i = 1,\dots,n.$$
$$\boldsymbol{w} \in \mathbb{R}^p_{\ge 0} \text{ and } \boldsymbol{w}^\top \mathbf{1} = 1.$$

We solve the above problem by appealing to a block coordinate descent as,

- Problem P_1: Fix $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0$, $\boldsymbol{w} = \boldsymbol{w}_0$, minimize $f(U, \boldsymbol{\Theta}_0, \boldsymbol{w}_0)$ w.r.t U.
- Problem P_2: Fix $U = U_0$, $\boldsymbol{w} = \boldsymbol{w}_0$, minimize $f(U_0, \boldsymbol{\Theta}, \boldsymbol{w}_0)$ w.r.t $\boldsymbol{\Theta}$.
- Problem P_3: Fix $U = U_0$, $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0$, minimize $f(U_0, \boldsymbol{\Theta}_0, \boldsymbol{w})$ w.r.t $\boldsymbol{w}$, subject to $\boldsymbol{w}^T\mathbf{1} = 1,\ w_i \ge 0$.

Problem, P_1, can be easily solved by assigning

$$u_{ij} = \begin{cases} 1, & j = \underset{1\le j\le k}{\operatorname{argmin}} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j\|^2_{\boldsymbol{w}^\beta} \\ 0, & \text{otherwise} \end{cases}$$

Problem P_2 can be easily solved by assigning, $\boldsymbol{\theta}_j = \frac{\sum_{i=1}^{n} u_{ij}\boldsymbol{x}_i}{\sum_{i=1}^{n} u_{ij}}$.

Problem P_3 can be solved using the *Frank-Wolfe* algorithm (Frank and Wolfe, 1956). Before proceeding further, let us formally write the problem as solving $\min_{\boldsymbol{w}\in\mathcal{C}}\ g(\boldsymbol{w})$. Here, $g(\boldsymbol{w}) = \sum_{\ell=1}^{p} w_\ell^\beta D_\ell + \Omega_{\boldsymbol{\lambda}}(\boldsymbol{w})$ and $\{\boldsymbol{w} : \boldsymbol{w}^T\mathbf{1} = 1, \boldsymbol{w} \ge \mathbf{0}\}$. It is easy to observe that both $g(\cdot)$ and the set $\mathcal{C}$ are convex. Though g is not differentiable, it is differentiable almost surely, and a sub-derivative of g can be taken as

$$\nabla g(\boldsymbol{w}) = \beta \boldsymbol{w}^{\beta-1} D_\ell + (\lambda_{\pi(1)}, \dots, \lambda_{\pi(p)})^\top.$$

Here π is the permutation which finds the increasing order of $\boldsymbol{w}$, i.e. π is such that $w_{\pi(\ell)} = w_{(\ell)}$. The mathematical simplicity of finding the $\operatorname{argmin}_{\boldsymbol{s}\in\mathcal{C}}\langle \boldsymbol{s}, \nabla g(\boldsymbol{w})\rangle$ makes it appropriate to use the algorithm as one can simply observe that as $\nabla g(\boldsymbol{w}) \in \mathbb{R}^p_{\ge 0}$, $\min_{\boldsymbol{s}\in\mathcal{C}}\langle \boldsymbol{s}, \nabla g(\boldsymbol{w})\rangle = \min_{\ell\in[p]} \nabla(g(\boldsymbol{w}))_\ell$. Note that for the Frank-Wolfe method, one needs to compute $\boldsymbol{s}$, which is $\boldsymbol{e}_k$, where $k = \operatorname{argmin}_{\ell\in[p]} \nabla(g(\boldsymbol{w}))_\ell$, and $\boldsymbol{e}_k$ is the i-th coordinate vector. The inner loop in Algorithm 1 highlights the Frank-Wolf steps for this particular problem. The results derived by Jaggi (2013) guarantee that the method converges in linear time (i.e., the estimates take at most $\mathcal{O}(1/\epsilon)$ for the solutions to reach an ϵ accuracy).

Note that the solution of the P_3 often lies in the boundary of the constraint set, to be precise at the corner of the convex bounded constraint set. When the solution of the *Frank-Wolfe* algorithm lies in the boundary, the algorithm guarantees its convergence at least at a linear rate. The convergence of the *Frank-Wolfe* algorithm thus determines the convergence of this clustering framework as each of the problems P_1 and P_2 are smooth and are solved by finding the critical values. If one runs the *Frank-Wolfe* until convergence, the cost function decreases monotonically. Hence the cost function, which is a function of three parameters, decreases monotonically and thus, converges by a simple application of the Bolzano–Weierstrass theorem as the cost is bounded below by 0.

4 Theoretical properties

In this section, we show that the convergence rates for the proposed OWL-k-means are significantly faster than its classical counterparts i.e., k-means and its variants. The complete proof of the result can be found in the supplement. We make the following standard assumption (Paul et al., 2021) on the data generation process.

Algorithm 1 OWL-k-means

Input: $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\} \in \mathbb{R}^{n\times p},\ \beta, k$
Output: $U, \Theta, \boldsymbol{w}$
Initialization: Randomly pick k data points $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k\}$ from $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$
while Objective (7) does not converge **do**
Step 1: Update U by

$$u_{ij}^{(t)} = \begin{cases} 1, & j = \underset{1\le j\le k}{\operatorname{argmin}} \|\boldsymbol{x}_i - \boldsymbol{\theta}_j^{(t-1)}\|_{\boldsymbol{w}^\beta}^2 \\ 0, & \text{otherwise} \end{cases}$$

Step 2: Update $\boldsymbol{\Theta}$ by $\boldsymbol{\theta}_j^{(t)} \leftarrow \frac{\sum_{i=1}^n u_{ij}^{(t)} \boldsymbol{x}_i}{\sum_{i=1}^n u_{ij}^{(t)}}$
Step 3: Update $\boldsymbol{w}$ by Frank-Wolfe method
Set $\boldsymbol{v}^{(0)} \leftarrow \boldsymbol{w}^{(t-1)}$
while $\|\boldsymbol{v}^{(\tau)} - \boldsymbol{v}^{(\tau+1)}\|_2 / \|\boldsymbol{v}^{(\tau)}\|_2 \ge \epsilon$ **do**
Find π is such that $v_{\pi(\ell)} = v_{(\ell)}$ by sorting.

$$m \leftarrow \underset{\ell\in[p]}{\operatorname{argmin}} \beta(v_\ell^{(\tau-1)})^{\beta-1} D_\ell + \lambda_{\pi(\ell)}$$

Set $\boldsymbol{s}^{(\tau-1)} \leftarrow \boldsymbol{e}_m$
Update $\boldsymbol{v}^{(\tau)} \leftarrow (1-\eta_t)\boldsymbol{v}^{(\tau-1)} + \eta_t \boldsymbol{s}^{(\tau-1)}$
end while
Set $\boldsymbol{w}^{(t)} \leftarrow \boldsymbol{v}^{(\tau)}$
end while

Assumption 1. *$\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are independent and identically distributed (i.i.d.) according to the distribution μ. Moreover, $\mu(B(M)) = 1$.*

The assumption ensures that the data are i.i.d. and bounded within the closed ball $B(M)$. For a fixed λ (after a scaling), the objective (6) can be cast as the following dual problem,

$$\min_{\boldsymbol{\Theta}\in\mathbb{R}^{k\times p}} \frac{1}{n}\sum_{i=1}^n \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\boldsymbol{x}_i - \boldsymbol{\theta}\|_{\boldsymbol{w}^\beta}^2, \text{ s.t. } \begin{cases} \Omega_{\boldsymbol{\lambda}}(\boldsymbol{w}) \le s \\ w_\ell \ge 0\ \forall \ell \in [p] \\ \boldsymbol{w}^\top \mathbf{1} = 1. \end{cases}$$

For simplicity of notations, let $\mathcal{W}_s = \{\boldsymbol{w} \in [0,1]^p : \Omega_{\boldsymbol{\lambda}}(\boldsymbol{w}) \le s \text{ and } \boldsymbol{w}^\top \mathbf{1} = 1.\}$. For notational simplicity, let $\hat{\mu}_n$ denote the empirical distribution of the data and $\varphi_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{x}) = \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\boldsymbol{x} - \boldsymbol{\theta}\|_{\boldsymbol{w}^\beta}^2$. Thus the objective is to solve the following problem,

$$\min_{\boldsymbol{\Theta}\in\mathbb{R}^{k\times p}, \boldsymbol{w}\in\mathcal{W}_s} \int \varphi_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{x}) d\hat{\mu}_n. \tag{8}$$

Let the solutions to the above empirical problem be, $\left(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{w}}\right)$, i.e.

$$\left(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{w}}\right) = \underset{\boldsymbol{\Theta}\in\mathbb{R}^{k\times p}, \boldsymbol{w}\in\mathcal{W}_s}{\operatorname{argmin}} \int \varphi_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{x}) d\hat{\mu}_n$$

Also let $(\boldsymbol{\Theta}^*, \boldsymbol{w}^*)$ be the solution to the population problem, i.e.

$$(\boldsymbol{\Theta}^*, \boldsymbol{w}^*) = \underset{\boldsymbol{\Theta}\in\mathbb{R}^{k\times p}, \boldsymbol{w}\in\mathcal{W}_s}{\operatorname{argmin}} \int \varphi_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{x}) d\mu$$

Our goal is to determine how close the solutions to the empirical problem are to the population problem by measuring the excess risk,

$$\begin{aligned} \mathfrak{R}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{w}}) &= \int \varphi_{\hat{\boldsymbol{\Theta}},\hat{\boldsymbol{w}}}(\boldsymbol{x}) d\mu - \inf_{\boldsymbol{\Theta}\in\mathbb{R}^{k\times p}, \boldsymbol{w}\in\mathcal{W}_s} \int \varphi_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{x}) d\mu \\ &= \int \varphi_{\hat{\boldsymbol{\Theta}},\hat{\boldsymbol{w}}}(\boldsymbol{x}) d\mu - \int \varphi_{\boldsymbol{\Theta}^*,\boldsymbol{w}^*}(\boldsymbol{x}) d\mu \end{aligned}$$

To bound this excess risk with high probability, we first observe that both $\hat{\boldsymbol{\Theta}}$ and $\boldsymbol{\Theta}^*$ lies in $B(M)$. This is formally stated in the following lemma.

Lemma 1. *Under Assumption 1, $\hat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}^* \in B(M)^k$.*

Thus, it is enough to restrict the search space to $B(M)$. We define the following function class

$$\mathscr{F} = \{\varphi_{\boldsymbol{\Theta},\boldsymbol{w}}(\cdot) : \boldsymbol{\Theta} \in B(M)^k,\ \boldsymbol{w} \in \mathcal{W}_s\}.$$

By appealing to Lemma 1, note that the excess risk can be bounded by the uniform concentration, $\sup_{f\in\mathscr{F}} \left|\int f d\hat{\mu}_n - \int f d\mu\right|$ as follows.

$$\begin{aligned} \mathfrak{R}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{w}}) &= \int \varphi_{\hat{\boldsymbol{\Theta}},\hat{\boldsymbol{w}}}(\boldsymbol{x}) d\mu - \int \varphi_{\boldsymbol{\Theta}^*,\boldsymbol{w}^*}(\boldsymbol{x}) d\mu \\ &\le 2 \sup_{f\in\mathscr{F}} \left| \int f d\hat{\mu}_n - \int f d\mu \right|. \end{aligned}$$

To bound this uniform deviation with high probability, we resort to finding suitable bounds on the Rademacher complexity of the function class $\mathscr{F}$, which is defined as follows:

Definition 1. *The population Rademacher complexity of a function class $\mathscr{F}$ is defined as follows,*

$$\mathcal{R}_n(\mathscr{F}) = \frac{1}{n}\mathbb{E} \sup_{f\in\mathscr{F}} \sum_{i=1}^n \epsilon_i f(\boldsymbol{X}_i).$$

Here, ϵ_i's are i.i.d. Rademacher random variables, i.e. $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = \frac{1}{2}$.

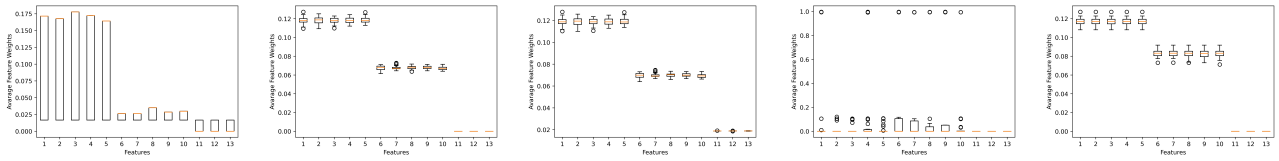
The following theorem derives a bound on the complexity of the function class $\mathscr{F}$. The notation $\lesssim$ hides all constants independent of n, p, k and s.

Theorem 1. *Under Assumption 1,*

$$\mathcal{R}_n(\mathcal{F}) \lesssim (s/\bar{\lambda} \vee 1)\sqrt{\frac{k \log p}{n}}.$$

If one closely looks at the proof of the result, one would observe that we do not need to use Dudley's chaining to bind the Rademacher complexity as opposed to Paul et al. (2021). We also show that the function class $\mathscr{F}$ is bounded. The following lemma asserts this claim.

Lemma 2. *Under Assumption 1, $\sup_{f\in\mathscr{F}} \|f\|_\infty \le 4M^2$.*



(a) EWP (b) LWK (c) Weighted-k-means (d) Sparse-k-means (e) OWL-k-means (ours)

Figure 2: The boxplot feature weights computed by our proposed algorithm and the peer algorithms over 30 simulations demonstrate the stability and efficacy of OWL-k-means.

Thus, appealing to Theorem 26.5 of Shalev-Shwartz and Ben-David (2014), we get the following bound on the excess risk by bounding the uniform concentration, $\sup_{f\in\mathscr{F}}\left|\int f d\hat{\mu}_n - \int f d\mu\right|$.

Theorem 2. *Grant Assumption 1. With probability at least* $1-\delta$,

$$\mathfrak{R}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{w}}) \lesssim (s/\bar{\lambda} \vee 1)\sqrt{\frac{k\log p}{n}} + \sqrt{\frac{\log(2/\delta)}{n}}.$$

It is easily observed that $\mathbb{E}\mathfrak{R}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{w}}) \lesssim (s/\bar{\lambda} \vee 1)\sqrt{\frac{k\log p}{n}}$, which is much faster than the parametric rate of $\sqrt{\frac{kp}{n}}$ (Bartlett et al., 1998) and also can be obtained from the results of Paul et al. (2021) under our assumption. Thus, under the additional assumption of sparsity, i.e., assuming that the optimal solution for the unconstrained problem lies in $\mathcal{W}_s$, the expected excess risk scales as $\sqrt{\log p}$ instead of $\sqrt{p}$. The change in the rate is similar to that of the differences in rates for OLS and Lasso, where the latter's convergence rate is similar to the derived rate for OWL-k-means. This theoretically establishes the superior performance of the proposal on single-cell RNA-seq or microarray data, where model sparsity is believed to hold. Also note that in a classical setting, when p is fixed, then the excess risk is $\mathcal{O}_P(1/\sqrt{n})$, recovering the classical parametric guarantees. The result is stated in the following corollary.

Corollary 1. *Suppose p and k are kept fixed, and Assumption 1 holds. Then,* $\mathfrak{R}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{w}}) = \mathcal{O}_P(1/\sqrt{n})$.

5 A Simulation Study on Feature Weights

To demonstrate the feature weighting and feature selection characteristic of our proposed algorithm, we use a simulation procedure where the generated datasets contain noisy features and features with varying variances. On each of the datasets, we run the proposal along with the peer methods and inspect the feature weight vector to analyze the capability of the algorithms to identify features with high variance and eliminate the noisy features. Intuitively, we choose such a simulation design where the cluster centroids are close to the cluster means for regular clusters. Thus, the Fisher information of each feature in this mixture model is inversely proportional to the within-cluster variance for the corresponding feature. The goal is to understand if this phenomenon is reflected in the feature weights.

Each data is generated with 200 data points, with 100 in each cluster, and each point has 13 features. Let $\boldsymbol{X}_i = (X_1^{(i)}, \ldots, X_{13}^{(i)})$ be a random point in the i-th cluster, where $i = \{1, 2\}$. The *data* is simulated as follows

- $X_j^{(1)}$ are i.i.d from $\mathcal{N}(0, 10)$, $\forall j \in \{1, \ldots, 5\}$
- $X_j^{(1)}$ are i.i.d from $\mathcal{N}(0, 50)$, $\forall j \in \{6, \ldots, 10\}$
- $X_j^{(2)}$ are i.i.d from $\mathcal{N}(100, 10)$, $\forall j \in \{1, \ldots, 5\}$
- $X_j^{(2)}$ are i.i.d from $\mathcal{N}(100, 50)$, $\forall j \in \{6, \ldots, 10\}$
- $X_j^{(i)}$ are i.i.d from $\mathcal{N}(0, 1)$, $\forall j \in \{11, \ldots, 13\}$, $i \in \{1, 2\}$. These random variables are independent of the above random variables.

We compute the feature weights produced by the algorithms over 30 runs and display them as boxplots in Fig. 2 for better comparison. It can be clearly seen that the noisy feature indexed from 11 to 13 are getting zero feature weights by OWL-k-means for most iterations by the OWL norm that penalizes the feature weights with a higher within-cluster sum of squares. Note that this is not the case for Sparse-k-means, Weighted-k-means, and EWP. Again, we can see that the features with higher variance are also penalized compared to the features with lower variance since the features with higher variance do not provide much helpful information for clustering the data. The feature weight vector has high values corresponding to features with lower variance consistently over 30 runs when treated by our proposed algorithm, compared to the feature weight vector produced by the Sparse-k-means, Weighted-k-means, and EWP-k-means. Also, we can see that the feature weight produced by the LW-k-means has a high interquartile range compared to OWL-k-means. These simulated datasets provide a brief and good picture of the ability to perform critical characteristics like feature weighting and feature selection by OWL-k-means in comparison to the state-of-the-art.

Table 1: Information regarding the datasets. Here, n is the number of data points, k denotes the number of clusters and p is the number of features.

Data	Source	n	p	k
Leukemia	Gordon et al. (2002)	72	3571	2
Lymphoma	Alizadeh et al. (2000)	62	4026	3
Brain	Pomeroy et al. (2002)	42	5597	5
Lung	Bhattacharjee et al. (2001)	203	12600	2
NCI9	ASU	60	9712	9

6 Real Data Experiments

We now study the experimental performance of OWL-k-means against the state-of-the-art on a suite of real-world data. For comparing the performance of various algorithms on the same dataset, we use the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) and the Normalized Mutual Information (Vinh et al., 2009) between the ground truth and the partition returned by the clustering algorithms. For both ARI and NMI, a value of 1 represents no mismatch and a value of 0 indicates a complete mismatch from the ground truth

As competitors, we consider some of the most effective feature-weight-based clustering method such as LW-k-means (Chakraborty and Das, 2020), k-means (MacQueen, 1967), W-k-means (Huang et al., 2005), IF-HCT-PCA (Jin and Wang, 2016), Sparse-k- means (Witten and Tibshirani, 2010) and Entropy Weighted Power k-means (EWP) (Chakraborty et al., 2020). To make sure that each method is on an equal footing, each algorithm is started with the same set of randomly chosen centroids, and for each dataset, iterations are carried out until convergence. The average performance index (i.e., ARI and NMI) between the ground truth and the obtained partitions are reported for 20 independent reruns. All the methods are tuned according to their original papers or to the best ARI/NMI values when not available. For OWL-k-means, we choose λ according to equation (4) and take $q = 0.1$, as suggested by Bogdan et al. (2015).

Ideal instances of high-dimensional data, where $p \gg n$, are microarray gene-expression datasets. In a typical microarray dataset, there are thousands of different gene-expression levels but very few samples, making it very challenging to analyze the clustering of the data. Such data have a meager signal-to-noise ratio, and is notoriously difficult to find a proper clustering of these data. The details of these data sets are given in Table 1. The average performance in terms of ARI for the peer methods is reported in Table 5 along with its standard error. It can be easily seen that OWL-k-means consistently performs better than the state-of-the-art despite hardly any tuning.

6.1 Test of Statistical Significance

We have conducted the Wilcoxon Rank Sum test (Wilcoxon, 1992), a non-parametric alternative to the two-sample t-test, to test the statistical significance of the performances of our method over the others. For this test, we have considered 20 ARI and NMI values, corresponding to each rerun, for each data, and for each method. We test for the null hypothesis: *average ARI (NMI) of our method = average ARI (NMI) of the competing method*, against the alternative that they are unequal. We record the p-value for the significance, whether it is less than 5%. If the p-value is significant, we conclude that there is enough evidence against the null. For each of the datasets, we record if the peer algorithms perform significantly better, worse or more or less the same compared to our proposal. The total number of times that they perform better (L), worse (W) or same (T) is reported at the end of each table. It can be easily seen that our method performs significantly better than the peers on most of the datasets.

6.2 Case Study on the Lymphoma dataset

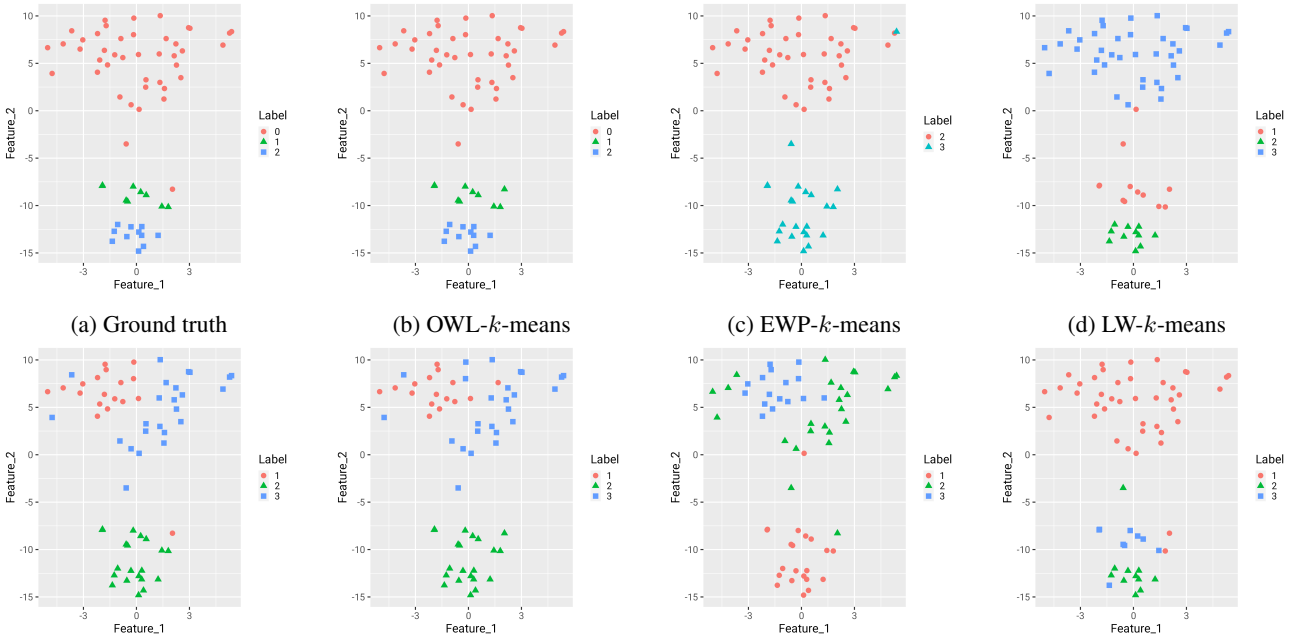
We evaluate the lymphoma dataset (Alizadeh et al., 2000), which consists of measurements of 4026 gene-expression levels, gathered across 62 samples. Out of the 62 samples, 42 are Diffuse Large B-Cell Lymphoma (DLBCL), 9 are Follicular Lymphoma (FL), and 11 are Chronic Lymphocytic Leukemia (CLL) cell samples. We compare the OWL-k- means to other baseline and cutting-edge clustering methods using this dataset to demonstrate its efficacy. We follow the same experimental protocols as before. Table 5 already suggests that OWL-k-means performs better than baseline k-means, W-k-means, and cutting-edge LW-k-means, Sparse-k-means and IF-HCT-PCA clustering algorithms. To better visualize the performance of different methods, we use t-SNE (Van der Maaten and Hinton, 2008) to reduce the dataset to two dimensions. From Fig. 3 it is evident that OWL-k-means resembles the ground truth compared to the state-of-the-art and further contributes to demonstrating its efficacy.

6.3 Single cell RNA-seq data

Understanding complex biological systems require analysis of expressions and regulations within each cell. With bulk analyses, gene expression is averaged across cells, whereas single-cell sequencing reveals the gene expressions by individual cells, providing a much deeper view of cell-to-cell variation. We use the OWL-k-means framework to treat the challenging single-cell RNA-seq data clustering problem. These datasets have a huge number of gene expressions compared to a minimal number of cells due to the limited and expensive nature of such sequencing and the low quality of cells/genes. There is also an issue of huge sparsity among single-cell objects. We test our clustering method's

Table 2: Average ARIs values for different algorithms on microarray gene expression datasets.(W/T/L : Win/ Tie/ Loss, †: results are significantly different as a result of Wilcoxon Rank-Sum test, ≈: results are statistically comparable as a result of Wilcoxon Rank-Sum test)

Dataset	Algorithms								
	k-means	W-k-means	Sparse-k-means	LW-k-means	EWP	IF-PCA-HCT	BP-k-means	WBMS	OWL-k-means
Leukemia	$0.53 \pm 0.38^{\dagger}$	$0.18 \pm 0.00^{\dagger}$	$0.73 \pm 0.00^{\dagger}$	$0.67 \pm 0.39^{\dagger}$	$0.21 \pm 0.00^{\dagger}$	$0.74 \pm 0.00^{\dagger}$	$0.84 \pm 0.00^{\dagger}$	$0.01 \pm 0.00^{\dagger}$	$\mathbf{0.89 \pm 0.00}$
Lymphoma	$0.53 \pm 0.23^{\dagger}$	$0.55 \pm 0.22^{\dagger}$	$0.41 \pm 0.00^{\dagger}$	$0.74 \pm 0.27^{\dagger}$	$0.52 \pm 0.17^{\dagger}$	$0.82 \pm 0.06^{\dagger}$	$0.94 \pm 0.00^{\dagger}$	$0.94 \pm 0.00^{\dagger}$	$\mathbf{0.95 \pm 0.00}$
Brain	$0.26 \pm 0.07^{\dagger}$	$0.45 \pm 0.02^{\dagger}$	$0.47 \pm 0.06^{\dagger}$	$0.32 \pm 0.10^{\dagger}$	$0.12 \pm 0.00^{\dagger}$	$0.48 \pm 0.03^{\dagger}$	$0.52 \pm 0.00^{\approx}$	$0.34 \pm 0.00^{\dagger}$	$\mathbf{0.52 \pm 0.08}$
Lung	$0.17 \pm 0.04^{\dagger}$	$0.01 \pm 0.00^{\dagger}$	$0.18 \pm 0.00^{\dagger}$	$\mathbf{0.25 \pm 0.00}^{\approx}$	$0.00 \pm 0.00^{\dagger}$	$\mathbf{0.25 \pm 0.00}^{\approx}$	$0.01 \pm 0.00^{\dagger}$	$0.01 \pm 0.00^{\dagger}$	$\mathbf{0.25 \pm 0.00}$
NCI9	$0.11 \pm 0.03^{\dagger}$	$0.16 \pm 0.03^{\dagger}$	$0.13 \pm 0.03^{\dagger}$	$0.11 \pm 0.04^{\dagger}$	$0.01 \pm 0.00^{\dagger}$	$0.04 \pm 0.01^{\dagger}$	$0.15 \pm 0.00^{\dagger}$	$0.10 \pm 0.00^{\dagger}$	$\mathbf{0.18 \pm 0.01}$
W/T/L	5/0/0	5/0/0	5/0/0	4/1/0	5/0/0	4/1/0	4/1/0	5/0/0	



(a) Ground truth (b) OWL-k-means (c) EWP-k-means (d) LW-k-means

(e) k-means (f) W-k-means (g) Sparse-k-means (h) IF-PCA-HCT

Figure 3: t-SNE plots for the Lymphoma dataset, showing the performance of OWL-k-means compared to other peer algorithms.

feature selection and weighting characteristics against specialized methods for clustering such datasets. Till now, the SC3 (Kiselev et al., 2017) clustering framework has, by far, been the most tested and well-performing algorithm that works best against these challenging single-cell datasets. We compare our algorithm against the clustering methods which are specially designed to tackle single-cell RNA-seq datasets, like, k-means-stable (Peyvandipour et al., 2020), Seuratv4 (Hao et al., 2021), RaceID3 (Grün, 2020), Monocle2 (Qiu et al., 2017), SC3 (Kiselev et al., 2017) and SOUP (Zhu et al., 2019). We use two such datasets: Biase (Biase et al., 2014) and Mouse Pancreas (Baron et al., 2016), to demonstrate the characteristic properties of our proposed clustering method. At first, we filtered the data by removing those genes expressed in less than 3 cells and then denoised the data using Deep Count Autoencoders (DCA) (Eraslan et al., 2019). After data cleaning, we follow the same experimental protocol as the previous sections to perform clustering, and the results in terms of the ARI values are tabulated in Table 3.

As we can see, the *Biase* dataset with 56 cells and 25737 genes gives a mean ARI 0.97, while testing the same for the *Mouse pancreas* data which contains 1886 cells and 14878 genes, the mean ARI 0.606 is the highest when compared with the state-of-the-art algorithms. Hence, Table 3 for average ARI values shows that our proposed algorithm performs significantly better over the peer algorithms. Comparative results in terms of the average NMI and t-SNE plots for visualization is included in the supplement for space economy.

7 Conclusions

This paper utilizes the power of OWL norms to lead to a novel clustering framework that draws good intuition from

Table 3: Average ARIs values for different algorithms on single-cell RNA-seq datasets.(W/T/L : Win/ Tie/ Loss, †: results are significantly different as a result of Wilcoxon Rank-Sum test, ≈: results are statistically comparable as a result of Wilcoxon Rank-Sum test)

Dataset	Algorithms						
	k-means-stable	Seuratv4*	RaceID3*	Monocle2*	SC3	SOUP	OWL-k-means (Ours)
Biase	$0.87 \pm 0.22^{\dagger}$	$0.59^{\dagger}$	$0.76^{\dagger}$	$0.69^{\dagger}$	$0.94 \pm 0.00^{\dagger}$	$0.86 \pm 0.00^{\dagger}$	$\mathbf{0.97 \pm 0.02}$
Mouse Pancreas	$0.42 \pm 0.03^{\dagger}$	$0.57^{\dagger}$	$0.32^{\dagger}$	$0.41^{\dagger}$	$0.43 \pm 0.02^{\dagger}$	$0.48 \pm 0.06^{\dagger}$	$\mathbf{0.606 \pm 0.091}$
W/T/L	2/0/0	2/0/0	2/0/0	2/0/0	2/0/0	2/0/0	

* : These algorithms are deterministic. Thus, their performance does not change over the runs.

classic and recent developments both in clustering and regression literature. The proposed objective promotes feature selection in an interpretable way in a center-based clustering framework with feature weighting. With an emphasis on simple updates, we derive an elegant combination of block-coordinate descent and the Frank-Wolfe algorithm to (locally) minimize the proposed objective. The paper also bridges the gap between the theory and practice of high-dimensional clustering by proving that the proposed OWL-k-means can achieve a Lasso-like fast error rate of $\mathcal{O}\left(\sqrt{k \log p/n}\right)$ under model sparsity, that was previously *not* observed in clustering literature. Our empirical studies show that OWL-k-means consistently outperforms its comparable variants. At the same time, its efficacy is thoroughly verified on a suite of high-dimensional microarray gene expression and single-cell RNA-seq data.

Several important research directions, however, remain open. A thorough theoretical analysis of the optimization scheme is lacking, and the model-selection properties of such estimates are yet unknown. Future work may explore this direction to seek explicit connections between the proposed method and its regression counterparts in a centroid-based setting and other clustering frameworks, such as convex or hierarchical clustering scenarios.

Code Availability

All the codes are available at: https://github.com/sayanpaul123/OWL_K_Means/.

References

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769):503–511.

Bao, R., Gu, B., and Huang, H. (2020). Fast oscar and owl regression via safe screening rules. In *International Conference on Machine Learning*, pages 653–663. PMLR.

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M., Melton, D. A., and Yanai, I. (2016). A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst*, 3(4):346–360.

Bartlett, P., Linder, T., and Lugosi, G. (1998). The minimax distortion redundancy in empirical quantizer design. *IEEE Transactions on Information Theory*, 44(5):1802–1813.

Bellman, R. (2003). *Dynamic Programming*. Dover Books on Computer Science Series. Dover Publications.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "nearest neighbor" meaningful? In *International conference on database theory*, pages 217–235. Springer.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., et al. (2001). Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences*, 98(24):13790–13795.

Biase, F. H., Cao, X., and Zhong, S. (2014). Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell rna sequencing. *Genome Res*, 24(11):1787–1796.

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). SLOPE—Adaptive variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103 – 1140.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123.

Chakraborty, S. and Das, S. (2020). Detecting meaningful clusters from high-dimensional data: A strongly consistent sparse center-based clustering approach. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence.

Chakraborty, S., Paul, D., Das, S., and Xu, J. (2020). Entropy weighted power k-means clustering. In *International conference on artificial intelligence and statistics*, pages 691–701. PMLR.

Chan, E. Y., Ching, W. K., Ng, M. K., and Huang, J. Z. (2004). An optimization algorithm for clustering using weighted dissimilarity measures. *Pattern recognition*, 37(5):943–952.

Chen, X., Ye, Y., Xu, X., and Huang, J. Z. (2012). A feature group weighting method for subspace clustering of high-dimensional data. *Pattern Recognition*, 45(1):434–446.

Cristofari, A., De Santis, M., Lucidi, S., and Rinaldi, F. (2017). New active-set frank-wolfe variants for minimization over the simplex and the ℓ_1-ball. *arXiv preprint arXiv:1703.07761*.

de Amorim, R. C. (2016). A survey on feature weighting based k-means algorithms. *Journal of Classification*, 33(2):210–242.

De Amorim, R. C. and Mirkin, B. (2012). Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering. *Pattern Recognition*, 45(3):1061–1075.

DeSarbo, W. S., Carroll, J. D., Clark, L. A., and Green, P. E. (1984). Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables. *Psychometrika*, 49(1):57–78.

Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S., and Theis, F. J. (2019). Single-cell rna-seq denoising using a deep count autoencoder. *Nature Communications*, 10(1):390.

Fraley, C. and Raftery, A. E. (1998). How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588.

Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110.

Gordon, G. J., Jensen, R. V., Hsiao, L.-L., Gullans, S. R., Blumenstock, J. E., Ramaswamy, S., Richards, W. G., Sugarbaker, D. J., and Bueno, R. (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research*, 62(17):4963–4967.

Grün, D. (2020). Revealing dynamics of gene expression variability in cell state space. *Nat Methods*, 17(1):45–49.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M. r., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P., and Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.

Huang, J. Z., Ng, M. K., Rong, H., and Li, Z. (2005). Automated variable weighting in k-means type clustering. *IEEE transactions on pattern analysis and machine intelligence*, 27(5):657–668.

Huang, J. Z., Xu, J., Ng, M., and Ye, Y. (2007). Weighting method for feature selection in k-means. In *Computational Methods of feature selection*, pages 209–226. Chapman and Hall/CRC.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Jaggi, M. (2013). Revisiting frank-wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435. PMLR.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

Jin, J. and Wang, W. (2016). Influential features pca for high dimensional clustering. *The Annals of Statistics*, 44(6):2323–2359.

Jing, L., Ng, M. K., and Huang, J. Z. (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on knowledge and data engineering*, 19(8):1026–1041.

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., and Hemberg, M. (2017). Sc3: consensus clustering of single-cell rna-seq data. *Nature Methods*, 14(5):483–486.

Kulis, B. and Jordan, M. I. (2011). Revisiting k-means: New algorithms via bayesian nonparametrics. *arXiv preprint arXiv:1111.0352*.

Kunisch, K. and Walter, D. (2021). On fast convergence rates for generalized conditional gradient methods with backtracking stepsize. *arXiv preprint arXiv:2109.15217*.

Lacoste-Julien, S. and Jaggi, M. (2015). On the global linear convergence of frank-wolfe optimization variants. *Advances in neural information processing systems*, 28.

Li, C. and Yu, J. (2006). A novel fuzzy c-means clustering algorithm. In *International Conference on Rough Sets and Knowledge Technology*, pages 510–515. Springer.

Liu, H. and Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4):491–502.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

MacQueen, J. (1967). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

Maurer, A. (2016). A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer.

McLachlan, G. J. and Rathnayake, S. (2014). On the number of components in a gaussian mixture model. *WIREs Data Mining and Knowledge Discovery*, 4(5):341–355.

McNicholas, P. D. (2016). Model-based clustering. *Journal of Classification*, 33(3):331–373.

Modha, D. S. and Spangler, W. S. (2003). Feature weighting in k-means clustering. *Machine learning*, 52(3):217–237.

Paul, D., Chakraborty, S., Das, S., and Xu, J. (2021). Uniform concentration bounds toward a unified framework for robust clustering. *Advances in Neural Information Processing Systems*, 34:8307–8319.

Peyvandipour, A., Shafi, A., Saberian, N., and Draghici, S. (2020). Identification of cell types from single cell data using stable clustering. *Scientific Reports*, 10(1):12349.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y., Goumnerova, L. C., Black, P. M., Lau, C., et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442.

Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A., and Trapnell, C. (2017). Single-cell mrna quantification and differential analysis with census. *Nature Methods*, 14(3):309–315.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Tsai, C.-Y. and Chiu, C.-C. (2008). Developing a feature weight self-adjustment mechanism for a k-means clustering algorithm. *Computational statistics & data analysis*, 52(10):4658–4672.

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Vinh, N. X., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wettschereck, D., Aha, D. W., and Mohri, T. (1997). A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, 11(1):273–314.

Wilcoxon, F. (1992). *Individual comparisons by ranking methods*. Springer.

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726.

Zhang, Z., Lange, K., and Xu, J. (2020). Simple and scalable sparse k-means clustering via feature ranking. *Advances in Neural Information Processing Systems*, 33:10148–10160.

Zhu, L., Lei, J., Klei, L., Devlin, B., and Roeder, K. (2019). Semisoft clustering of single-cell data. *Proceedings of the National Academy of Sciences*, 116(2):466–471.

Appendix

A Proofs from Section 4

Lemma 3. *Let* $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$. *Then,* $\min_{1\le j\le p} a_j - \min_{1\le j\le p} b_j \le \|\boldsymbol{a}-\boldsymbol{b}\|_2$.

Proof. Suppose $j^* \in \operatorname{argmin} b_j$. Then,

$$\begin{aligned}\min_{1\le j\le p} a_j - \min_{1\le j\le p} b_j &= \min_{1\le j\le p} a_j - b_{j^*} \\ &\le a_{j^*} - b_{j^*} \le \|\boldsymbol{a}-\boldsymbol{b}\|_2.\end{aligned}$$

□

A.1 Proof of Lemma 1

Proof. Let $\operatorname{Proj}^{\boldsymbol{w}}_{B(M)}(\boldsymbol{a})$ denote the projection of $\boldsymbol{a}$ onto $B(M)$ w.r.t. the $\|\cdot\|_{\boldsymbol{w}}$-norm. For any $\boldsymbol{v} \in B(M)$, using the obtuse angle property, we obtain, $\langle \boldsymbol{\theta} - \operatorname{Proj}^{\boldsymbol{w}}_{B(M)}(\boldsymbol{\theta}), \boldsymbol{v} - \operatorname{Proj}^{\boldsymbol{w}}_{B(M)}(\boldsymbol{\theta})\rangle_{\boldsymbol{w}} \le 0$ due to the convexity of $B(M)$. Let, $\boldsymbol{x} \in B(M)$, then,

$$\begin{aligned}&\|\boldsymbol{x}-\boldsymbol{\theta}\|^2_{\boldsymbol{w}} \\ =&\|\boldsymbol{x} - \operatorname{Proj}^{\boldsymbol{w}}_{B(M)}(\boldsymbol{\theta})\|^2_{\boldsymbol{w}} + \|\operatorname{Proj}^{\boldsymbol{w}}_{B(M)}(\boldsymbol{\theta}) - \boldsymbol{\theta}\|^2_{\boldsymbol{w}} \\ &- 2\langle \boldsymbol{\theta} - \operatorname{Proj}^{\boldsymbol{w}}_{B(M)}(\boldsymbol{\theta}), \boldsymbol{x} - \operatorname{Proj}^{\boldsymbol{w}}_{B(M)}(\boldsymbol{\theta})\rangle_{\boldsymbol{w}} \\ \ge&\|\boldsymbol{x} - \operatorname{Proj}^{\boldsymbol{w}}_{B(M)}(\boldsymbol{\theta})\|^2_{\boldsymbol{w}} + \|\operatorname{Proj}^{\boldsymbol{w}}_{B(M)}(\boldsymbol{\theta}) - \operatorname{Proj}^{\boldsymbol{w}}_{B(M)}\|^2_{\boldsymbol{w}} \\ \ge&\|\boldsymbol{x} - \operatorname{Proj}^{\boldsymbol{w}}_{B(M)}(\boldsymbol{\theta})\|^2_{\boldsymbol{w}}\end{aligned}$$

Thus,

$$\int \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\boldsymbol{x}-\boldsymbol{\theta}\|^2_{\boldsymbol{w}} dP(\boldsymbol{x}) \ge \int \|\boldsymbol{x} - \operatorname{Proj}^{\boldsymbol{w}}_{B(M)}(\boldsymbol{\theta})\|^2_{\boldsymbol{w}} dP(\boldsymbol{x}).$$

This further implies that, $\min_{\boldsymbol{\Theta}\in\mathbb{R}^{k\times p},\boldsymbol{w}} \int \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\boldsymbol{x} - \boldsymbol{\theta}\|^2_{\boldsymbol{w}} dP(\boldsymbol{x}) \ge \min_{\boldsymbol{\Theta}\in B(M)^k,\boldsymbol{w}} \int \|\boldsymbol{x}-\boldsymbol{\theta}\|^2_{\boldsymbol{w}} dP(\boldsymbol{x})$, from which, we conclude that $\boldsymbol{\Theta}^*(P) \in B(M)$. □

A.2 Proof of Theorem 1

Proof. Suppose $\xi_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{x}) = (\|\boldsymbol{x}-\boldsymbol{\theta}_1\|^2_{\boldsymbol{w}}, \ldots, \|\boldsymbol{x}-\boldsymbol{\theta}_k\|^2_{\boldsymbol{w}})$. Suppose $\{\epsilon_i\}_{i\in[n]}$ and $\{\sigma_{ij}\}_{i,j\in[n]}$ be two sets of independent Rademacher random variables. Note that,

$$\begin{aligned}\varphi_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{x}) - \varphi_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{y}) &= \min_{j\in[k]}(\xi_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{x}))_j - \min_{j\in[k]}(\xi_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{y}))_j \\ &\le \|\xi_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{x}) - \xi_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{y})\|_2\end{aligned}$$

Here the last inequality follows from Lemma 3. Thus appealing to inequality (1) of Maurer (2016), we get,

$$\begin{aligned}&\mathbb{E}\sup_{\boldsymbol{\Theta}\in B(M)^k,\boldsymbol{w}\in\mathcal{W}} \sum_{i=1}^n \epsilon_i \varphi_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{X}_i) \\ \le&\sqrt{2}\mathbb{E}\sup_{\boldsymbol{\Theta}\in B(M)^k,\boldsymbol{w}\in\mathcal{W}} \sum_{i=1}^n\sum_{j=1}^k \sigma_{ij}(\xi_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{X}_j))_j \\ =&\sqrt{2}\mathbb{E}\sup_{\boldsymbol{\Theta}\in B(M)^k,\boldsymbol{w}\in\mathcal{W}} \sum_{i=1}^n\sum_{j=1}^k \sigma_{ij}\|\boldsymbol{X}_i - \boldsymbol{\theta}_j\|^2_{\boldsymbol{w}^\beta} \\ \le&\sqrt{2}\mathbb{E}\sup_{\boldsymbol{w}\in\mathcal{W}} \sum_{i=1}^n\sum_{j=1}^k \sigma_{ij}\|\boldsymbol{X}_i\|^2_{\boldsymbol{w}^\beta} \qquad (9) \\ &+ 2\sqrt{2}\mathbb{E}\sup_{\boldsymbol{\Theta}\in B(M)^k,\boldsymbol{w}\in\mathcal{W}} \sum_{i=1}^n\sum_{j=1}^k \sigma_{ij}\langle \boldsymbol{X}_i, \boldsymbol{\theta}_j\circ\boldsymbol{w}^\beta\rangle \qquad (10) \\ &+ \sqrt{2}\mathbb{E}\sup_{\boldsymbol{\Theta}\in B(M)^k,\boldsymbol{w}\in\mathcal{W}} \sum_{i=1}^n\sum_{j=1}^k \sigma_{ij}\|\boldsymbol{\theta}_j\|^2_{\boldsymbol{w}^\beta} \qquad (11)\end{aligned}$$

We bound the terms individually as follows.

First term

$$\mathbb{E}\sup_{\boldsymbol{w}\in\mathcal{W}} \sum_{i=1}^n\sum_{j=1}^k \sigma_{ij}\langle \boldsymbol{w}^\beta, \boldsymbol{X}_i^2\rangle \le M^2\mathbb{E}\left|\sum_{i=1}^n\sum_{j=1}^k \sigma_{ij}\right| \le M^2\sqrt{kn}.$$

Here the last inequality can be seen as a simple application of Jensen's inequality.

Second term

$$\begin{aligned}&\mathbb{E}\sup_{\boldsymbol{\Theta}\in B(M)^k,\boldsymbol{w}\in\mathcal{W}} \sum_{i=1}^n\sum_{j=1}^k \sigma_{ij}\langle \boldsymbol{X}_i, \boldsymbol{\theta}_j\circ\boldsymbol{w}^\beta\rangle \\ \le&\mathbb{E}\sup_{\boldsymbol{\Theta}\in B(M)^k,\boldsymbol{w}\in\mathcal{W}} \sum_{j=1}^k \langle \sum_{i=1}^n \sigma_{ij}\boldsymbol{X}_i, \boldsymbol{\theta}_j\circ\boldsymbol{w}^\beta\rangle \\ \le&\mathbb{E}\sup_{\boldsymbol{\Theta}\in B(M)^k,\boldsymbol{w}\in\mathcal{W}} \sum_{j=1}^k \|\sum_{i=1}^n \sigma_{ij}\boldsymbol{X}_i\|_\infty \|\boldsymbol{\theta}_j\circ\boldsymbol{w}^\beta\|_1 \\ \le&k\mathbb{E}\sup_{\boldsymbol{\Theta}\in B(M)^k,\boldsymbol{w}\in\mathcal{W}} \|\sum_{i=1}^n \sigma_{ij}\boldsymbol{X}_i\|_\infty \|\boldsymbol{\theta}_j\circ\boldsymbol{w}^\beta\|_1 \\ \le&kM(s/\bar\lambda\vee 1)\mathbb{E}\|\sum_{i=1}^n \epsilon_i\boldsymbol{X}_i\|_\infty \\ \le&kM(s/\bar\lambda\vee 1)\mathbb{E}\|\sum_{i=1}^n \epsilon_i\boldsymbol{X}_i\|_\infty \\ =&kM(s/\bar\lambda\vee 1)\mathbb{E}\sup_{\ell\in[p]}|\sum_{i=1}^n \epsilon_i X_{il}| \\ \le&\sqrt{2\log p}kM(s/\bar\lambda\vee 1)\mathbb{E}|\sum_{i=1}^n \epsilon_i X_{il}| \\ \le&\sqrt{2\log p}kM^2(s/\bar\lambda\vee 1)\sqrt{n}\end{aligned}$$

Third term One can bound the third term by following the recipe of bounding the first terms.

$$\mathbb{E}\sup_{\boldsymbol{\Theta}\in B(M)^k,\boldsymbol{w}\in\mathcal{W}} \sum_{i=1}^n\sum_{j=1}^k \sigma_{ij}\|\boldsymbol{\theta}_j\|^2_{\boldsymbol{w}^\beta} \le M^2(s/\bar\lambda\vee 1)\sqrt{kn}.$$

Using the above bounds in equation (11), we get the desired result. □

Lemma 2 $\sup_{\boldsymbol{\Theta}\in B(M)^{k\times p},\, \boldsymbol{w}\in[0,1]^p} \|\varphi_{\boldsymbol{\Theta},\boldsymbol{w}}\|_\infty \leq 4M^2$.

Proof. We observe that, for any $\boldsymbol{x}, \boldsymbol{\Theta} \in B(M)^k$, $\boldsymbol{w} \in [0,1]^p$,

$$\begin{aligned} 0 \leq \varphi_{\boldsymbol{\Theta},\boldsymbol{w}}(\boldsymbol{x}) \leq & \|\boldsymbol{x}-\boldsymbol{\theta}_1\|^2_{\boldsymbol{w}^2} \\ \leq & 2\left(\|\boldsymbol{x}\|^2_{\boldsymbol{w}^2} + \|\boldsymbol{\theta}_1\|^2_{\boldsymbol{w}^2}\right) \\ \leq & 2\|\boldsymbol{w}\|_\infty \left(\|\boldsymbol{x}\|_2^2 + \|\boldsymbol{\theta}_1\|_2^2\right) \leq 4M^2. \end{aligned}$$

□

A.3 Proof of Theorem 2

Proof. We note that

$$\begin{aligned} & \mathfrak{R}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{w}}) \\ = & \int \varphi_{\hat{\boldsymbol{\Theta}},\hat{\boldsymbol{w}}}(\boldsymbol{x}) d\mu - \int \varphi_{\boldsymbol{\Theta}^*,\boldsymbol{w}^*}(\boldsymbol{x}) d\mu \\ = & \int \varphi_{\hat{\boldsymbol{\Theta}},\hat{\boldsymbol{w}}}(\boldsymbol{x}) d\mu - \int \varphi_{\hat{\boldsymbol{\Theta}},\hat{\boldsymbol{w}}}(\boldsymbol{x}) d\hat{\mu}_n \\ & + \int \varphi_{\hat{\boldsymbol{\Theta}},\hat{\boldsymbol{w}}}(\boldsymbol{x}) d\hat{\mu}_n - \int \varphi_{\boldsymbol{\Theta}^*,\boldsymbol{w}^*}(\boldsymbol{x}) d\hat{\mu}_n \\ & + \int \varphi_{\boldsymbol{\Theta}^*,\boldsymbol{w}^*}(\boldsymbol{x}) d\hat{\mu}_n \int \varphi_{\boldsymbol{\Theta}^*,\boldsymbol{w}^*}(\boldsymbol{x}) d\mu \\ \leq & 2 \sup_{f\in\mathscr{F}} \left| \int f d\hat{\mu}_n - \int f d\mu \right|. \end{aligned}$$

The theorem now follows from appealing to Theorem 1. □

Corollary 1

Proof. We note from Theorem 2 that, with probability at least $1-\delta$, $\mathfrak{R}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{w}}) \lesssim \frac{1}{\sqrt{n}}$. Thus, for any fixed δ, $\sqrt{n}\mathfrak{R}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{w}}) \lesssim 1$, which implies that $\sqrt{n}\mathfrak{R}(\hat{\boldsymbol{\Theta}}, \hat{\boldsymbol{w}})$ is tight. □

B Tables and Plots for section 6

C Social Impact

Our work focuses on algorithmic and theoretical contributions to unsupervised learning of high-dimensional data. There are no immediate privacy or ethical concerns, but by addressing the persistent problem of presence of a huge number of noisy features, broader impacts extend beyond methodological contributions when the interpretation of pattern discoveries from the output of unsupervised learning methods have wider implications. Clustering has been used for countless applications, including community detection, drug discovery, and gene identification for cancers and other diseases. In such settings where the interpretations and decisions based on clustering solutions have significant scientific and societal bearing, it is critical that the outliers are not mistaken as original data while solving for optimal solutions or baseline truth.

That said, we have been careful not to overstate our claims. While theoretical and empirical evidence supports that we can significantly reduce the effect of noisy features, users should not view our method as a panacea for the problem. Our algorithm provides only a partial remedy to a long-standing challenge faced by clustering methods, and we emphasize it may eliminate some but not all biases that may affect interpretations and decisions based on solutions output by unsupervised algorithms.

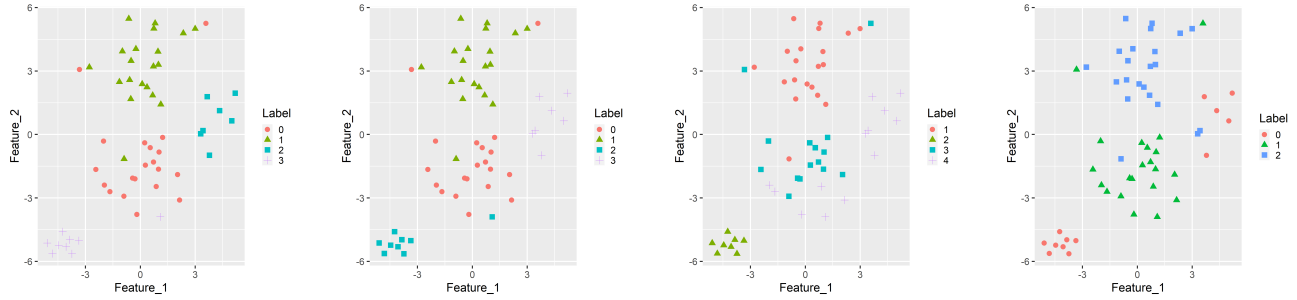
Table 4: Average NMIs values for different algorithms on single-cell RNA-seq datasets.(W/T/L : Win/ Tie/ Loss, †: results are significantly different as a result of Wilcoxon Rank-Sum test, ≈: results are statistically comparable as a result of Wilcoxon Rank-Sum test)

Dataset	Algorithms						
	k-means-stable	Seuratv4*	RaceID3*	Monocle2*	SC3	SOUP	OWL-k-means (Ours)
Biase	$0.89 \pm 0.16^{\dagger}$	$0.69^{\dagger}$	$0.82^{\dagger}$	$0.79^{\dagger}$	$0.93 \pm 0.00^{\dagger}$	$0.85 \pm 0.00^{\dagger}$	$\mathbf{0.97 \pm 0.02}$
Mouse Pancreas	$0.54 \pm 0.07^{\dagger}$	$0.75^{\dagger}$	$0.62^{\dagger}$	$0.70^{\dagger}$	$0.73 \pm 0.01^{\dagger}$	$0.71 \pm 0.01^{\dagger}$	$\mathbf{0.76 \pm 0.02}$
W/T/L	2/0/0	2/0/0	2/0/0	2/0/0	2/0/0	2/0/0	

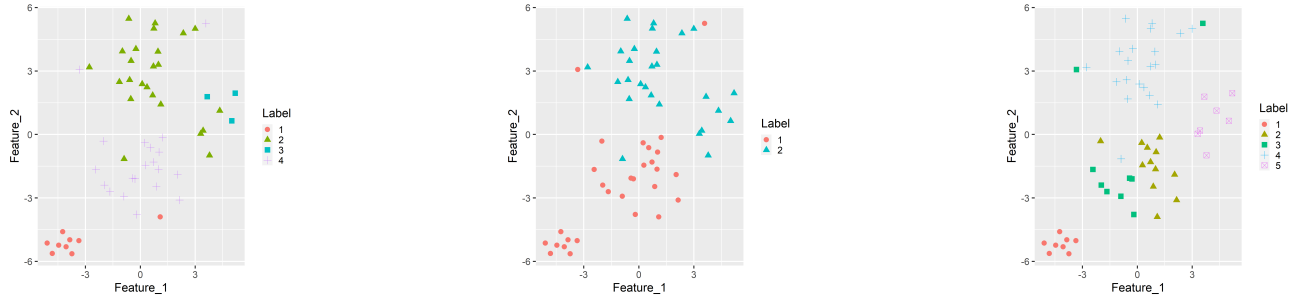
* : These algorithms are deterministic. Thus, their performance does not change over the runs.

Table 5: Average NMIs values for different algorithms on microarray gene expression datasets.(W/T/L : Win/ Tie/ Loss, †: results are significantly different as a result of Wilcoxon Rank-Sum test, ≈: results are statistically comparable as a result of Wilcoxon Rank-Sum test)

Dataset	Algorithms								
	k-means	W-k-means	Sparse-k-means	LW-k-means	EWP	IF-PCA-HCT	BP-k-means	WBMS	OWL-k-means
Leukemia	$0.53 \pm 0.30^{\dagger}$	$0.46 \pm 0.00^{\dagger}$	$0.62 \pm 0.00^{\dagger}$	$0.63 \pm 0.33^{\dagger}$	$0.22 \pm 0.00^{\dagger}$	$0.67 \pm 0.00^{\dagger}$	$0.74 \pm 0.00^{\dagger}$	$0.05 \pm 0.00^{\dagger}$	$\mathbf{0.81 \pm 0.00}$
Lymphoma	$0.66 \pm 0.15^{\dagger}$	$0.39 \pm 0.22^{\dagger}$	$0.58 \pm 0.00^{\dagger}$	$0.80 \pm 0.17^{\dagger}$	$0.60 \pm 0.03^{\dagger}$	$0.72 \pm 0.06^{\dagger}$	$0.92 \pm 0.00^{\dagger}$	$0.91 \pm 0.00^{\dagger}$	$\mathbf{0.93 \pm 0.00}$
Brain	$0.43 \pm 0.05^{\dagger}$	$0.58 \pm 0.06^{\dagger}$	$0.59 \pm 0.03^{\dagger}$	$0.49 \pm 0.07^{\dagger}$	$0.19 \pm 0.00^{\dagger}$	$0.55 \pm 0.03^{\dagger}$	$0.64 \pm 0.00^{\dagger}$	$0.47 \pm 0.00^{\dagger}$	$\mathbf{0.64 \pm 0.07}$
Lung	$0.11 \pm 0.08^{\dagger}$	$0.01 \pm 0.00^{\dagger}$	$0.20 \pm 0.00^{\dagger}$	$\mathbf{0.27 \pm 0.00}^{\approx}$	$0.01 \pm 0.00^{\dagger}$	$\mathbf{0.27 \pm 0.00}^{\approx}$	$0.00 \pm 0.00^{\dagger}$	$0.02 \pm 0.00^{\dagger}$	$\mathbf{0.27 \pm 0.00}$
NCI9	$0.38 \pm 0.02^{\dagger}$	$0.39 \pm 0.12^{\dagger}$	$0.40 \pm 0.03^{\dagger}$	$0.41 \pm 0.04^{\dagger}$	$0.05 \pm 0.00^{\dagger}$	$0.37 \pm 0.03^{\dagger}$	$0.44 \pm 0.00^{\dagger}$	$0.36 \pm 0.00^{\dagger}$	$\mathbf{0.46 \pm 0.02}$
W/T/L	5/0/0	5/0/0	5/0/0	4/1/0	5/0/0	4/1/0	5/0/0	5/0/0	



(a) Ground truth (b) OWL-k-means (c) SC3 (d) k-means-stable



(e) SOUP (f) Seurat (g) RaceID3

Figure 4: t-SNE plots for the Biase dataset, showing the performance of OWL-k-means compared to other peer algorithms for the single-cell RNA seq data.
Note: The t-SNE plot for Monocle cannot be procured due to the internal issues of the Monocle package which was updated recently, we can assure to reproduce the same at the time of submission if accepted.