Clustering by Sum of Norms: Stochastic Incremental Algorithm, Convergence and Cluster Recovery

Ashkan Panahi [1] **Devdatt Dubhashi** [2] **Fredrik D. Johansson** [3] **Chiranjib Bhattacharyya** [4]

Abstract

Standard clustering methods such as K–means, Gaussian mixture models, and hierarchical clustering, are beset by local minima, which are sometimes drastically suboptimal. Moreover the number of clusters K must be known in advance. The recently introduced sum–of–norms (SON) or Clusterpath convex relaxation of k-means and hierarchical clustering shrinks cluster centroids toward one another and ensure a unique global minimizer. We give a scalable stochastic incremental algorithm based on proximal iterations to solve the SON problem with convergence guarantees. We also show that the algorithm recovers clusters under quite general conditions which have a similar form to the unifying *proximity condition* introduced in the approximation algorithms community (that covers paradigm cases such as Gaussian mixtures and planted partition models). We give experimental results to confirm that our algorithm scales much better than previous methods while producing clusters of comparable quality.

1. Introduction

Clustering is perhaps the most fundamental problem in unsupervised learning. Many clustering algorithms have been proposed in the literature (Jain et al., 1999), including K-means, spectral clustering, Gaussian mixture models and hierarchical clustering, to solve problems with respect to a wide range of cluster shapes. However, much research has pointed out that these methods all suffer from instabilities. For example, the formulation of K-means is NP-hard and the typical way to solve it is the Lloyds method, which requires randomly initializing the clusters. However, one needs to know the number of clusters in advance and different initializations may lead to significantly different final cluster results.

Lindsten et al. (2011) and Hocking et al. (2011) proposed the following convex optimization procedure for clustering, called SON ("Sum of norms" clustering) by the former and Clusterpath by the latter;

$$\min_{\{u_i \in \mathbb{R}^m\}} \frac{1}{2} \sum_{i=1}^{n} \|x_i - u_i\|_2^2 + \lambda \sum_{i<j} \|u_i - u_j\|_2 \quad (1)$$

The main idea of the formulation is that if input data points x_i and x_j belong to the same cluster, then their corresponding centroids u_i and u_j should be forced to be the same. Intuitively, this is due to the fact that the second term is a regularization term that enforces zeroes in the vector consisting of entries $\|u_i - u_j\|$ and can be seen as a generalization of the fused Lasso penalty. From another point of view, the regularization term can be seen as an $\ell_{1,2}$ norm, i.e., the sum of ℓ_2 norms. Such a *group norm* is known to encourage block sparse solutions (Bach et al., 2012). Thus for many pairs (i, j), we expect to enforce $u_i = u_j$.

Lindsten et al. (2011) used an off–the–shelf convex solver, CVX to generate solution paths. Hocking et al. (2011) introduced three distinct algorithms for the three most commonly encountered norms. For the ℓ_1 norm, the objective function separates, and they solve the convex clustering problem by the exact path following method designed for the fused lasso. For the ℓ_1 and ℓ_2 norms, they employ subgradient descent in conjunction with active sets. Recently, Chi & Lange (2015); Chen et al. (2015) introduce two similar generic frameworks for minimizing the convex clustering objective function with an arbitrary norm. One approach solves the problem by the alternating direction method of multipliers (ADMM), while the other solves it by the alternating minimization algorithm (AMA). However both algorithms have issues with scalablity.

Moreover, none of these papers provide any theoretical guarantees about the cluster recovery property of the algorithm. The first theoretical result on cluster recovery was shown by Zhu et al. (2010): if the samples are drawn from

[1]ECE, North Carolina State University, Raleigh, NC [2]CSE, Chalmers University of Technology, Göteborg, Sweden [3]IMES, MIT, Cambridge, MA [4]CSA, IISc, Bangalore, India. Correspondence to: Ashkan Panahi <panahi1986@gmail.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

two cubes, each being one cluster, then SON can provably recover the clusters provided that the distance between the two cubes is larger than a threshold which depends (linearly) on the size of the cube and the ratio of numbers of samples in each cluster. Unfortunately, the conditions for recovery represent an extremely narrow special case: only two clusters which both have to be cubes. Moreover in their paper, there is no algorithm or analysis of the speed of convergence. No other theoretical guarantees for SON are known previously.

Here we develop a new algorithm in the spirit of recent advances in stochastic methods for large scale optimization (Bottou et al., 2016) to solve the optimization problem (1). We give a convergence analysis and provide quite general cluster recovery guarantees.

There has been a flurry of advances (Johnson & Zhang, 2013; Defazio et al., 2014; Schmidt et al., 2016) in developing algorithms for solving optimization problems for the case when the objective consists of the sum of two convex functions: one is the average of a large number of smooth component functions, and the other is a general convex function that admits a proximal mapping (and the whole objective function is strongly convex). The optimization (1) is of this form but here we exploit the structure of (1) further by observing that the second function can also be split into component functions. This results in an incremental algorithm with proximal iterations consisting of very simple and natural steps. Our algorithm can be seen as a special case of the methods of Bertsekas (2011). We compute the proximal operator in closed form to yield very simple and cheap iterations. Using the fact that the proximal operator is non-expansive, we refine and strengthen Bertsekas' convergence results. The stochastic incremental nature of our algorithm makes it highly suited to large scale problems (Bottou et al., 2016) in contrast to the methods in Chi & Lange (2015); Chen et al. (2015).

We show that the SON formulation (1) provides strong cluster recovery properties that go far beyond the special case considered in Zhu et al. (2010). Our cluster recovery conditions are similar in spirit to the unifying general conditions recently formulated in A. Kumar (2010); P.Awasthi (2012) of the form that the means of the clusters are well–separated, i.e., the distance between the means of any two clusters is at least $\Omega(k)$ standard deviations (the notion of standard deviations is based on the spectral norm of the matrix whose rows represent the difference between a point and the mean of the cluster to which it belongs). Besides containing the result of Zhu et al. (2010) as a special case, the condition essentially recovers the well known cluster recovery conditions for paradigm examples such as mixtures of Gaussians and planted partition models. The algorithms in A. Kumar (2010); P.Awasthi (2012) are based on an SVD-based initialization followed by applying Lloyd's K–means algorithm, so K must be known in advance. Our method does not need to know K and is independent of any initialization.

A summary of our contributions are:

- We develop a new incremental proximal algorithm for the SON optimization problem (1).
- We give a convergence analysis for our algorithm that refines and strengthens the analysis in Bertsekas (2011).
- We show that the SON formulation (1) provides strong cluster recovery guarantees that is far more general than previously known recovery results, essentially similar to the recently discovered unifying center separation conditions.
- We give experimental results giving evidence that our algorithm produces clusters of comparable quality to previous methods but scales much better to large scale problems.

2. Related Work

The SON formulation first appeared in (Lindsten et al., 2011) and in closely related forms in Hocking et al. (Hocking et al., 2011). Lindsten et al (Lindsten et al., 2011) used an off–the–shelf convex solver, CVX to generate solution paths. Hocking et al. (Hocking et al., 2011) introduced three distinct algorithms for the three most commonly encountered norms. For the ℓ_1 norm, the objective function separates, and they solve the convex clustering problem by the exact path following method designed for the fused lasso. For the ℓ_1 and ℓ_2 norms, they employ subgradient descent in conjunction with active sets. Neither provides any theoretical results on cluster recovery. Chi et al (Chi & Lange, 2015; Chen et al., 2015) introduce two similar generic frameworks for minimizing the convex clustering objective function with an arbitrary norm. One approach solves the problem by the alternating direction method of multipliers (ADMM), while the other solves it by the alternating minimization algorithm (AMA). The first (and only) theoretical results on cluster recovery are in (Zhu et al., 2010) but this is a very simple special case of exactly two cube shaped clusters that are well separated. This work also does not develop a specialized algorithm for the SON formulation.

3. Cluster Recovery

To express our results, we first review few definitions:

Definition 1. Take a finite set $X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^m$ and its partitioning $\mathcal{V} = \{V_1, V_2, \ldots, V_K\}$, where each

V_k is a subset of X. We say that a map ϕ on X *perfectly recovers* $\mathcal{V}$ when $\phi(x_i) = \phi(x_j)$ is equivalent to x_i, x_j belonging to the same cluster, or in other words, there exist distinct vectors $v_1, v_2, \ldots, v_K$ such that $\phi(x_i) = v_\alpha$ holds whenever $x_i \in V_\alpha$.

Definition 2. For any set $S \subset \mathbb{R}^m$, its diameter is defined as

$$D(S) = \sup\{\|x - y\|_2 \mid x, y \in S\}.$$

Moreover, for any finite set $T \subset \mathbb{R}^m$ we define its separation as

$$d(T) = \min\{\|x - y\|_2 \mid x, y \in S, x \neq y\}$$

and its *Euclidean centroid* as

$$c(T) = \frac{\sum\limits_{x \in T} x}{|V|}.$$

Finally, for any family of mutually disjoint finite sets $\mathcal{T} = \{T_i \subset \mathbb{R}^m\}$, we define $\mathcal{C}(\mathcal{T}) = \{c(T_i)\}$.

Definition 3. Take a finite set $X = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^m$ and its partitioning $\mathcal{V} = \{V_1, V_2, \ldots, V_K\}$. We call a partitioning $\mathcal{W} = \{W_1, W_2, \ldots, W_L\}$ of X a *coarsening* of $\mathcal{V}$ if each partition W_l is obtained by taking the union of a number of partitions V_k. Further, $\mathcal{W}$ is called the trivial coarsening of $\mathcal{V}$ if $\mathcal{W}$ has exactly one element, i.e. $\mathcal{W} = \{X\}$. Otherwise, it is called a non-trivial coarsening.

Based on the above definitions, our result can be explained as follows:

Theorem 1. *Consider a finite set $X = \{x_i \in \mathbb{R}^m \mid i = 1, 2, ..., n\}$ of vectors and its partitioning $\mathcal{V} = \{V_1, V_2, \ldots, V_K\}$. Take the SON optimization in (1). Denote its optimal solution by $\{\bar{u}_i\}$ and define the map $\phi : x_i \to \phi(x_i) = \bar{u}_i$.*

1. *If*
$$\max_{V \in \mathcal{V}} \frac{D(V)}{|V|} \leq \lambda \leq \frac{d(\mathcal{C}(\mathcal{V}))}{2n\sqrt{K}},$$
then the map ϕ perfectly recovers $\mathcal{V}$.

2. *If*
$$\max_{V \in \mathcal{V}} \frac{D(V)}{|V|} \leq \lambda \leq \max_{V \in \mathcal{V}} \frac{\|c(X) - c(V)\|_2}{|X| - |V|},$$
then the map ϕ perfectly recovers a non-trivial coarsening of $\mathcal{V}$.

Proof. We introduce associated centroid optimization:

$$\min_{\{v_\alpha \in \mathbb{R}^m\}} \frac{1}{2} \sum_{i=1}^{K} \|v_\alpha - c_\alpha\|_2^2 n_\alpha + \lambda \sum_{\alpha \neq \beta} n_\alpha n_\beta \|c_\alpha - c_\alpha\|_2 \tag{2}$$

where

$$c_\alpha = \frac{\sum\limits_{i \in V_\alpha} x_i}{n_\alpha}$$

We prove the following results, which clearly imply Theorem 1:

1. Suppose that for every $\alpha \in [K]$,
$$\frac{\max\limits_{i,j \in V_\alpha} \|x_i - x_j\|}{n_\alpha} \leq \lambda.$$
Then, $u_i = v_\alpha$ for $i \in V_\alpha$ is a global solution of the SON clustering.

2. If all c_αs are distinct and $\frac{d}{2n\sqrt{K}} \geq \lambda$ where $d = \min\limits_{\alpha \neq \beta} \|c_\alpha - c_\beta\|$, then all centroids v_α are distinct.

3. If $\max\limits_{\alpha} \frac{\|c_\alpha - c\|}{n - n_\alpha} \geq \lambda$ where $c = \sum\limits_{i=1}^{n} x_i/n$, then at least two centroids v_α are distinct.

To prove the above, notice that the solution of the centroid optimization satisfies

$$c_\alpha - v_\alpha = \lambda \sum_\beta n_\beta z_{\alpha,\beta}$$

where $\|z_{\alpha,\beta}\| \leq 1$, $z_{\alpha,\beta} = -z_{\beta,\alpha}$ and whenever $v_\alpha \neq v_\beta$, the relation $z_{\alpha,\beta} = \frac{v_\alpha - v_\beta}{\|v_\alpha - v_\beta\|_2}$ holds. Now, for the solution $u_i = v_\alpha$ for $i \in V_\alpha$, define

$$z'_{ij} = \begin{cases} z_{\alpha,\beta} & \alpha \neq \beta \\ \frac{x_i - x_j}{\lambda n_\alpha} & \alpha = \beta \end{cases},$$

where $i \in V_\alpha$, $j \in V_\beta$. It is easy to see that $\|z'_{ij}\|_2 \leq 1$, $z'_{ij} = -z'_{ji}$ and whenever $u_i \neq u_j$, we have that $z'_{ij} = \frac{u_i - u_j}{\|u_i - u_j\|_2}$. Further for each i,

$$\lambda \sum_j z'_{i,j} = \lambda \sum_\beta z_{\alpha,\beta} n_\beta + \sum_{j \in V_\alpha} \frac{x_i - x_j}{n_\alpha}$$

$$= c_\alpha - v_\alpha + x_i - c_\alpha = x_i - v_\alpha = x_i - u_i$$

This shows that the local optimality conditions for the SON optimization holds and proves item a.

For item b, denote the solution of the centroid optimization by $v_\alpha(\lambda)$ and notice that the solution of SON consists of distinct elements $v_\alpha = c_\alpha$ and is continuous at $\lambda = 0$. Hence, v_α:s remain distinct in an interval $\lambda \in [0,\ \lambda_1)$. Take λ_0 as the supremum of all possible λ_1:s. Hence, the solution in $\lambda \in [0,\ \lambda_0)$ contains distinct element and at $\lambda = \lambda_0$ contains two equal elements (otherwise, one can extend $[0,\ \lambda_0)$ to some $[0,\ \lambda_0 + \epsilon)$, which is against λ being supremum). Now, notice that for $\lambda \in [0\ \lambda_0)$ the objective

function is smooth at the optimal point. Hence, $v_\alpha(\lambda)$ is differentiable and satisfies

$$\delta = \left[\frac{\mathrm{d}v_\alpha}{\mathrm{d}\lambda}\right]_\alpha = H^{-1}\frac{\partial g}{\partial \lambda} \tag{3}$$

where $[.]_\alpha$ and $[.]_{\alpha,\beta}$ denote block vectors and block matrices respectively. Moreover, H and g are the Hessian and the gradient of the objective function at the optimal point. It is possible, by explicitly expanding H and g, to show that $\|\delta\|_2 \leq n\sqrt{K}$ (see the supplementary material for more detailed derivations).

Hence,

$$\left\|\frac{\mathrm{d}v_\alpha}{\mathrm{d}\lambda}\right\|_2 \leq \|\delta\|_2 \leq \sqrt{K}n$$

This yields for $\lambda < \lambda_0$ to

$$\|v_\alpha(\lambda) - v_\beta(\lambda)\|_2 = \left\|c_\alpha - c_\beta + \int_0^\lambda \left(\frac{\mathrm{d}v_\alpha}{\mathrm{d}\lambda} - \frac{\mathrm{d}v_\beta}{\mathrm{d}\lambda}\right)\mathrm{d}\lambda\right\|_2$$

$$\geq \|c_\alpha - c_\beta\|_2 - \int_0^\lambda \left\|\frac{\mathrm{d}v_\alpha}{\mathrm{d}\lambda} - \frac{\mathrm{d}v_\beta}{\mathrm{d}\lambda}\right\|_2 \mathrm{d}\lambda$$

$$\geq d - 2n\lambda\sqrt{K}$$

Since at $\lambda = \lambda_0$, we have that $v_\alpha = v_\beta$ for some $\alpha \neq \beta$, we get that $d - 2n\lambda_0\sqrt{K} \leq 0$ or $\lambda_0 \geq d/2n\sqrt{K}$. this proves item b.

For item c, Take a value of λ, where $v_1 = v_2 = \ldots = v_K$. It is simple to see that in this case $v_\alpha = c$. The optimality condition leads to

$$c - c_\alpha = \lambda \sum_{\beta \neq \alpha} z_{\alpha,\beta} n_\beta$$

Hence, $\|c - c_\alpha\|_2 \leq \lambda(n - n_\alpha)$. This proves item c. □

Remark 1. The study in Zhu et al. (2010) establishes some results for the special case of two clusters in rectangular boxes. In this special case, we observe that our result improves theirs.

Proof. Consider the notation in Zhu et al. (2010) with two clusters V_1, V_2 and notice that $\lambda = \alpha/2$ (α denotes regularization parameter in Zhu et al. (2010)). Moreover, $D(V_i) \leq \|\mathbf{s}_i\|$ as $\|\mathbf{s}_i\|$ is the diameter of the rectangle surrounding V_i. We observe that

$$\frac{w_{1,2}}{n} \geq D(V_i)\frac{2n_{3-i}\left(\frac{n_i-1}{n_i^2}\right)+1}{n_{3-i}+n_i} \geq \frac{D(V_i)}{n_i}$$

for $i = 1, 2$, which shows that the condition $\lambda \geq \frac{w_{1,2}}{n}$ in Zhu et al. (2010) is tighter than $\lambda \geq \max D(V)/|V|$ in ours. On the other hand,

$$\begin{aligned}\frac{\|c(V_i) - c(X)\|}{n - n_i} &= \frac{\|c(V_i) - \frac{c(V_1)n_1 + c(V_2)n_2}{n_1+n_2}\|}{n - n_i} \\ &= \frac{\|c(V_1) - c(V_2)\|_2}{n} \\ &= \frac{d(\mathcal{C}(\mathcal{V}))}{n}\end{aligned} \tag{4}$$

Hence, the condition $\lambda \leq \frac{d(\mathcal{C}(\mathcal{V}))}{n}$ in Zhu et al. (2010) is the same as our condition $\lambda \leq \frac{\|c(V_i)-c(X)\|}{n-n_i}$. □

Remark 2. The second result in Theorem 1 reflects a hierarchical structure in the SON clusters: Under weaker condition than the first part, SON may merge some clusters and provide larger clusters than the true ones. In a recursive way, SON clustering can be applied to each of these large clusters to refine them, which improves the guarantees in Theorem 1. We postpone careful study of this method to future work.

3.1. Comparison with Center Separation Conditions

Recently, there have been a number of theoretical results of the form that if we have data points generated by a mixture of K probability distributions, then one can cluster the data points into the K clusters, one corresponding to each component, provided the means of the different components are well–separated. There are different notions of well-separated, but mainly, the (best known) results can be qualitatively stated as: "If the means of every pair of densities are at least poly(K) standard deviations apart, then we can learn the mixture in polynomial time.". These results generally make heavy use of the generative model and particular properties of the distributions (Indeed, many of them specialize to Gaussians or independent Bernoulli trials).

Kumar and Kannan (A. Kumar, 2010) and Awasthi and Sheffet (P.Awasthi, 2012) unified these into a general deterministic condition which can be roughly stated as follows: "If the means of every pair of clusters are at least $\Omega(K)$ times standard deviations apart, then we can learn the mixture in polynomial time." Here the spectral norm of the matrix $A - C$ scaled by $\frac{1}{\sqrt{n}}$ plays the role of standard deviation, where A is the data matrix and C is the matrix of cluster centers. More formally, for any two distinct clusters α, β,

$$\|c(V_\alpha) - c(V_\beta)\|_2 \geq K\left(\frac{1}{\sqrt{n_\alpha}} + \frac{1}{\sqrt{n_\beta}}\right)\|A - C\| \tag{5}$$

Our condition is similar in spirit:

$$\|c(V_\alpha) - c(V_\beta)\|_2 \geq \sqrt{K}\left(\frac{n}{n_\alpha}d(V_\alpha)\right) \tag{6}$$

If $n_\alpha \geq wn$ for all clusters α, then this becomes

$$\|c(V_\alpha) - c(V_\beta)\|_2 \geq \frac{\sqrt{K}}{w} d(V_\alpha). \tag{7}$$

In the sequel, we specialize the above discussion in a number of examples and provide an explicit comparison of our result with the center separation condition. In some cases, our condition is slightly tighter than the center separation guarantees, but we remind that the latter is obtained by applying K-means and a SVD-based initialization, which can be intractable in large problems, while our techniques scales with the problem size more suitably.

3.1.1. Mixtures of Gaussians

Suppose we have a mixture of K Gaussians in d dimensions with mixture weights $w_1, \cdots, w_K$, let $w := \min_i w_i$ and let $\mu_1, \cdots, \mu_K$ denote their means respectively. If we have $n = \Omega(\text{poly}(d/w))$ points sampled from this mixture distribution, then with high probability, the center separation condition is satisfied if:

$$\|\mu_r - \mu_s\| \geq \frac{cK\sigma}{\sqrt{w}} \text{polylog}(d/w).$$

Here σ is the maximum variance in any direction of any of the Gaussians. Our cluster recovery condition (7) is satisfied if:

$$\|\mu_r - \mu_s\| \geq \frac{cK\sigma}{w} \text{polylog}(n).$$

3.1.2. Planted Partition Model

In the planted partition model of McSherry, a set of n points is implicitly partitioned into K groups. There is an (unknown) $K \times K$ matrix of probabilities P . We are given a graph G on these n points, where an edge between two vertices from groups r and s is present with probability $P_{r,s}$. We can consider these n points $x_1, \cdots, x_n \in R^n$ where coordinate j in x_i is 1 if $(i,j) \in G$ and 0 otherwise. The center μ_r of cluster r has in coordinate j the value $P_{r,\psi(j)}$, where $\psi(j)$ is the cluster vertex j belongs to. Kumar and Kannan show that the center separation condition holds with probability at least $1 - \delta$ if:

$$\|\mu_r - \mu_s\| \geq c\sigma^2 K(\frac{1}{w} + \log\frac{n}{\delta})$$

where c is a large constant, w is such that every group has size at least $w \cdot n$ and $\sigma^2 := \max_{r,s} P_{r,s}$. Our center separation condition (7) is satisfied if:

$$\|\mu_r - \mu_s\| \geq c\frac{\sigma^2 K}{w}\sqrt{n}$$

3.1.3. Regular and Directed Clusters

Besides the stochastic models, we take a closer look at the result in A. Kumar (2010) and identify deterministic cases where the SON has better performance than the proved bounds for K-means. These cases essentially guarantee that the term $\|A-C\|$ in (5) remains large and the bound therein becomes highly restrictive:

Definition 4. We say that a partition $\mathcal{V} = \{V_1, V_2, \ldots, V_K\}$ of $X = \{x_1, x_2, \ldots, x_n\}$ is $(\delta, \gamma)-$expanded if

$$|\{x \in V \mid \|x - c(V)\|_2 \geq \delta\}| \geq \gamma|V|.$$

We further say that this partition is (w, D, ϵ, γ)-regular if for all $V \in \mathcal{V}$ we have $D(V) \geq D$, $|V| \geq wn$ and it is $(\epsilon D, \gamma)-$expanded.

Definition 5. We say that a set $X = \{x_1, x_2, \ldots, x_n\}$ is $\theta-$directed if there exists a unit vector $v \in \mathbb{R}^m$ such that

$$\sum_{x \in X \setminus \{c(X)\}} \frac{|v^T(x - c(X))|^2}{\|x - c(X)\|_2^2} \geq \theta|X|$$

For a (w, D, ϵ, γ)-regular partition, the bound in (5) implies $d(\mathcal{C}(\mathcal{V})) \geq \frac{2cK\epsilon D\sqrt{\gamma n}}{\sqrt{mwn}}$. This is because

$$\begin{aligned} \|A - C\|^2 = \sigma_{\max}\left(\sum_{i=1}^n \delta_i\delta_i^T\right) &\geq \frac{\text{Tr}\left(\sum_{i=1}^n \delta_i\delta_i^T\right)}{m} \\ = \frac{\sum_{i=1}^n \|\delta_i\|_2^2}{m} &= \gamma\epsilon^2 D^2 \frac{n}{m} \end{aligned} \tag{8}$$

where $\delta_i = x_i - c(V_\alpha)$ for $i \in V_\alpha$. Notice that our conditions can be implied by $d(\mathcal{C}(\mathcal{V})) \geq \frac{2nD\sqrt{K}}{wn}$. Hence,SON can improve K-means if $m \leq wKc^2\gamma\epsilon^2$, which means that the number of clusters K is large and the smallest fraction of cluster size w is $\Omega(1)$.

If the (w, D, ϵ, γ)-regular partition is further $\theta-$directed we may improve the previous bounds as

$$\sigma_{\max}\left(\sum_{i=1}^n \delta_i\delta_i^T\right) \geq \sum_{x \in X} |v^T\delta_i|^2 \geq \gamma\epsilon^2 D^2 n\theta$$

Hence (5) implies $d(\mathcal{C}(\mathcal{V})) \geq \frac{2cK\epsilon D\sqrt{\gamma\theta n}}{\sqrt{wn}}$. This means that SON improves K-means if $wK \geq \frac{1}{c^2\epsilon^2\gamma\theta}$, i.e. the number of clusters is higher than a fixed value.

4. Stochastic Splitting Algorithm

Our implementation is identical to the so-called proximal-based incremental technique in Bertsekas (2011), which is

performed in a way that it requires little amount of calculations (precisely $O(m)$ and independent of other parameters) in each iteration. The proximal-based incremental method is a variant of the stochastic gradient technique, in problems where many terms in the objective function are not differentiable, and the local gradient steps are replaced by local proximal operators. To perform the proximal-based incremental method, we first write the SON objective function as

$$\Phi(u_1, u_2, \ldots, u_n) = \sum_{i<j} \phi_{ij}(u_i, u_j)$$

where

$$\phi_{ij}(u_i, u_j) = \frac{1}{2n}\|x_i - u_i\|_2^2 + \frac{1}{2n}\|x_j - u_j\|_2^2 + \lambda\|u_i - u_j\|.$$

Then, we introduce and explicitly calculate the proximal operator $\Pi_{ij}^{(\mu)}$ of ϕ_{ij} with step size μ as

$$\begin{aligned}&\Pi_{ij}^{(\mu)}(u_i, u_j) = \\ &\arg\min_{u_i', u_j'} \phi_{ij}(u_i', u_j') + \tfrac{1}{2\mu}\|u_i' - u_i\|_2^2 + \tfrac{1}{2\mu}\|u_j' - u_j\|_2^2 \\ &= \mathcal{T}_{\lambda\mu}(u_i + \mu x_i, u_j + \mu x_j), \end{aligned} \tag{9}$$

where we also introduce the pairwise soft-thresholding operator $\mathcal{T}_\eta(y, z) =$

$$\begin{cases} \left(y + \eta\frac{z-y}{\|z-y\|_2}, z + \eta\frac{y-z}{\|y-z\|_2}\right) & \|y - z\| \geq 2\eta \\ \left(\frac{y+z}{2}, \frac{y+z}{2}\right) & \|y - z\| < 2\eta \end{cases}, \tag{10}$$

and the final equality is obtained by the local optimality conditions and straightforward calculations. Our algorithm simply consists in iteratively applying randomly selected proximal operators. This is depicted in Algorithm 1.

Algorithm 1 Stochastic Splitting Algorithm

Input: The data vectors $\{x_k\}_{k=1}^n$ and step sizes $\{\mu_k\}_{k=1}^\infty$
Initialization: Set $u_1, u_2, \ldots, u_n$ arbitrarily (we use $u_1 = u_2 = \ldots = u_n = 0$)
for $k = 1, 2, \ldots$ **do**
  Select a pair (i, j) with $i < j$ uniformly randomly.
  Update $(u_i, u_j) \leftarrow \Pi_{ij}^{(\mu_k)}(u_i, u_j)$
end for

4.1. Convergence Analysis

Convergence of proximal-based incremental method is discussed in Bertsekas (2011). We further elaborate on the convergence by further exploitation of the non-expansiveness property of proximal operators. This allows us to complement the result in Bertsekas (2011) in the following two directions: First, we establish convergence in the probability sense (uniform convergence), while the result in Bertsekas (2011) is pointwise. Second, we prove guaranteed speed of convergence with probability one. We present these results by the following theorem. In this theorem, we consider fixed data dimension m and bounded data vectors (i.e. $\|\mathbf{x}_k\| \leq C$ for some absolute constant C).

Theorem 2.

1. *Assume that $\{\mu_k\}$ is non-increasing $\sum_0^\infty \mu_k = \infty$ and $\sum_0^\infty \mu_k^2 < \infty$. Then, the sequence $\mathbf{U}_k$ converges to $\tilde{\mathbf{U}}$ in the following strong probability sense:*

$$\forall \epsilon > 0;\ \lim_{k\to\infty} \Pr\left(\sup_{l\geq k} \|\mathbf{U}_l - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 > \epsilon\right) = 0 \tag{11}$$

2. *Take $\mu_k = \frac{\mu_1}{k^\alpha}$ for $k = 1, 2, \ldots$ and $\frac{2}{3} < \alpha < 1$. For sufficiently small values of $\epsilon > 0$ the relation*

$$\|\mathbf{U}_l - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 = O\left(\frac{n^4}{l^{3\alpha-2-\epsilon}}\right)$$

holds for every l, n with probability 1.

Proof. We skip many steps in our proof for lack of space. These steps can be found in the supplement. Denote by $\mathbf{U}_k$ a matrix where the i^{th} column is the value of u_i at the k^{th} iteration. Define

$$\psi_\mu(\mathbf{U}) = \mathcal{E}\left(\mathbf{U}_{k+1} \mid \mathbf{U}_k = \mathbf{U}, \mu_k = \mu\right), \tag{12}$$

Starting from $\bar{\mathbf{U}}_0 = \mathbf{U}_0$ (the initialization of the algorithm), we define the characteristic sequence $\{\bar{\mathbf{U}}_k\}_{k=0}^\infty$ by the following iteration:

$$\bar{\mathbf{U}}_{k+1} = \psi_{\mu_k}(\bar{\mathbf{U}}_k)$$

Our proof is based on the following two results, which we prove in the supplementary material:

i We have that

$$\Pr\left(\sup_k \|\mathbf{U}_k - \bar{\mathbf{U}}_k\|_{\mathrm{F}}^2 + \sum_{l=k}^\infty \mu_l^2 > \lambda\right) \leq \frac{\sum_{k=0}^\infty \mu_k^2}{\lambda} \tag{13}$$

ii Define $\tilde{\mathbf{U}}$ as the unique optimal solution of the SON optimization and suppose that $\{\mu_k\}$ is a non-increasing sequence. There exists a universal constant a such that $\|\bar{\mathbf{U}}_k - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2$ is upper bounded by

$$a\sum_{l=0}^{k-1} \mu_l^2 e^{-\frac{2}{n^2}\sum_{s=l+1}^{k-1}\mu_s} + \|\mathbf{U}_0 - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 e^{-\frac{2}{n^2}\sum_{s=0}^{k-1}\mu_s}$$

To prove Theorem 2, define $\mathcal{U}^k = \{\bar{\mathbf{U}}_l^k\}_{l=0}^{\infty}$ as the sequence obtained by starting from $\bar{\mathbf{U}}_0^k = \mathbf{U}_k$ and applying

$$\bar{\mathbf{U}}_{l+1}^k = \psi_{\mu_{l+k}}(\bar{\mathbf{U}}_l^k)$$

Take arbitrary (non-zero) positive numbers ϵ, δ. Take λ such that $\lambda \geq \frac{2}{\delta}\sum_{l=0}^{\infty}\mu_l^2$. Take some values l_0, k which we specialize later. Now, we define two outcomes H_1 and H_2:

$$H_1 : \forall k \geq 0;\ \|\mathbf{U}_k - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 \leq \lambda$$

$$H_2 : \forall l \geq 0; \|\bar{\mathbf{U}}_l^k - \mathbf{U}_{l+k}\| \leq \frac{\epsilon}{4}$$

From item (i), it is simple to see that $\Pr(H_1^c)$ and $\Pr(H_2^c)$ are less than $\delta/2$. Furthermore we can show by (ii) that under $H_1 \cap H_2$ and suitable l_0, k:

$$\forall l > l_0;\ \|\mathbf{U}_{l+k} - \tilde{\mathbf{U}}\|_2^2 \leq 2(\|\mathbf{U}_{l+k} - \bar{\mathbf{U}}_l^k\|_{\mathrm{F}}^2 + \|\bar{\mathbf{U}}_l^k - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2)$$

$$\leq 2(\frac{\epsilon}{4} + \frac{\epsilon}{4}) = \epsilon$$

This is detailed in the supplement. We conclude that

$$\Pr(\sup_{l > l_0 + k} \|\mathbf{U}_l - \tilde{\mathbf{U}}\|_2^2 > \epsilon) \leq \Pr(H_1^c) + \Pr(H_2^c) \leq \delta$$

which proves part (1) of Theorem.

For part (2), define $k_r = r^{\gamma}$, $\lambda_r = r^{-\beta}$, where $\gamma = \frac{1-\frac{\epsilon}{2}}{1-\alpha}$, $\beta < \gamma(2\alpha - 1) - 1$, and the outcomes:

$$Q_r :\ \sup_{l \geq 0} \|\mathbf{U}_{l+k_r} - \bar{\mathbf{U}}_l^{k_r}\|_{\mathrm{F}}^2 > \lambda_r.$$

By item (i), we have that $\sum_{r=1}^{\infty} \Pr(Q_r) < \infty$. Hence by Borel-Cantelli lemma, $Q_{r_0}^c, Q_{r_0+1}^c, Q_{r_0+2}^c, \ldots$ simultaneously hold for some r_0 with probability 1. For simplicity and without loss of generality, we assume that $r_0 = 0$ as it does not affect the asymptotic rate. Then for any $r > 0$, we have that

$$\sup_{l \geq 0} \|\mathbf{U}_{l+k_r} - \bar{\mathbf{U}}_l^{k_r}\|_{\mathrm{F}}^2 \leq \lambda_r$$

In particular,

$$\|\mathbf{U}_{k_{r+1}} - \bar{\mathbf{U}}_{l_r}^{k_r}\|_{\mathrm{F}}^2 \leq \lambda_r$$

where $l_r = k_{r+1} - k_r$. From item (ii), we also conclude that

$$\|\bar{\mathbf{U}}_{l_r}^{k_r} - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 \leq A\sum_{t=0}^{l_r-1} \frac{1}{(t+k_r)^{2\alpha}} e^{-2a\sum_{s=t+1}^{l_r-1}\frac{1}{(s+k_r)^{\alpha}}}$$

$$+ \|\mathbf{U}_{k_r} - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 e^{-2a\sum_{s=0}^{l_r-1}\frac{1}{(s+k_r)^{\alpha}}}$$

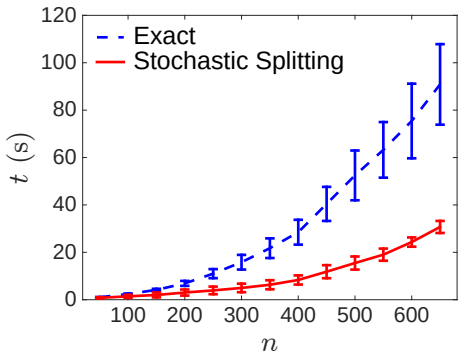


Figure 2. Running times of the exact SON clustering algorithm (implemented in CVX) and stochastic splitting for samples from a mixture of two Gaussians with increasing sample size.

where we introduce $\mu_1 = bn^2$ and $A = 4an^4b^2$ for simplicity. This leads to

$$\|\mathbf{U}_{k_{r+1}} - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 \leq Le^{Lk_r^{1-\alpha} - Lk_{r+1}^{1-\alpha}}\|\mathbf{U}_{k_r} - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 +$$

$$2\lambda_r + A\sum_{t=0}^{l_r} \frac{1}{(t+k_r)^{2\alpha}} e^{L(k_r+t)^{1-\alpha} - Lk_{r+1}^{1-\alpha}}$$

where L is a suitable constant with different values at different occurrences. Postponing few more steps to the supplementary material, we obtain that

$$\|\mathbf{U}_{k_r} - \tilde{\mathbf{U}}\|_{\mathrm{F}}^2 \leq \frac{L\log r}{r^{\beta - \frac{\epsilon}{2}}} \leq \frac{L}{r^{\beta-\epsilon}}$$

Take $k_r < l \leq k_{r+1}$. We obtain that

$$\|\mathbf{U}_l - \tilde{\mathbf{U}}\|_2^2 \leq 2(\|\mathbf{U}_{k_r} - \tilde{\mathbf{U}}\|_2^2 + \|\mathbf{U}_{k_r} - \mathbf{U}_l\|_2^2)$$

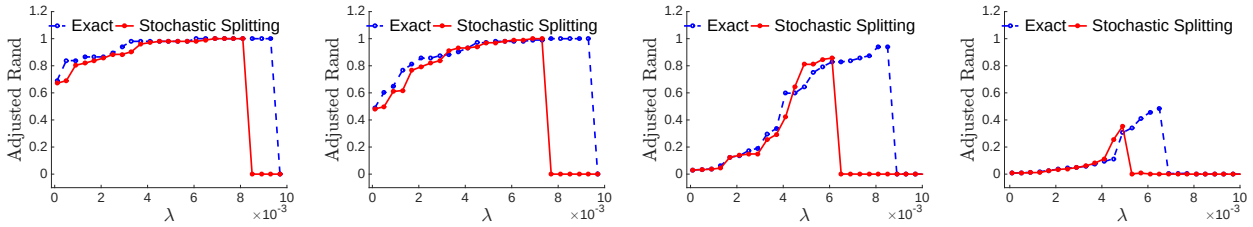
$$\leq 2\lambda_r + \frac{L}{r^{\beta-\epsilon}} \leq \frac{L}{r^{\beta-\epsilon}} \leq \frac{L}{l^{\frac{\beta-\epsilon}{\gamma}}}$$

By taking $\beta = \gamma(2\alpha - 1) - 1$, we obtain part (2).

□

5. Experiments

We evaluate the proposed stochastic splitting algorithm in the task of clustering points generated by Gaussians mixture models. We compare the results to the exact algorithm proposed by Lindsten et al. (2011) in terms of a) the quality of the produced clustering and b) the time spent solving the optimization problem. The results of both algorithms are dense embeddings of the points that are then thresholded to form clusters. The clusters are the largest subsets of nodes such that the maximum pairwise distance within the subset is less than τ. The stochastic splitting algorithm is implemented as in Algorithm 1. We observed in practice



(a) $\sigma^2 = 0.01$ (b) $\sigma^2 = 0.02$ (c) $\sigma^2 = 0.05$ (d) $\sigma^2 = 0.1$

Figure 1. Adjusted Rand index for different choices of λ. Each plot represents quality of the clustering produced by solving the SON objective exactly or with stochastic splitting. The different plots represent clustering of 200 samples from a mixture of two Gaussians in $\mathbb{R}^2$ with fixed separation $d = \sqrt{2}$ and variance σ^2.

that a heuristic for adaptively setting the step-size improved robustness and rate of convergence. Specifically, the step size was reduced by a constant factor whenever the average change in the objective over successive rounds in a small window was positive. If the same average was negative, but small in absolute value, the step size was increased by a small constant factor.

The data is generated from Gaussian mixture models with two components in $\mathbb{R}^2$ where the means are separated by $d = \sqrt{2}$ and the variance σ^2 is varied. The number of samples is also varied, to illustrate the computational gains of the stochastic splitting method. As pointed out by Lindsten et al. (2011), the choice of the regularization parameter λ is perhaps the most challenging hurdle in applying SON clustering. Choosing λ too high might result in a single large cluster, and choosing it too low may cause each point to be represented by its own cluster. While this problem is of great importance in applications, we focus on the relative performance of the Lindsten et al. (2011) algorithm (CVX) and stochastic splitting (SS). We report the adjusted Rand index (Rand, 1971) as measure of cluster quality, and would like to emphasize that this does not rely on identifying the number of clusters beforehand.

Results The results of the experiments are presented in Figures 1 and 2. We see in Figure 1 that the quality of the clustering produced by the stochastic splitting algorithm is comparable to that of the exact algorithm. This pattern is consistant across choices of σ, where a high σ implies low sample separation between the clusters. We also note that the range of λ for which the stochastic splitting algorithm achieves as good result as the exact algorithm is less wide than for the exact. We believe this is due to the stochastic nature of the algorithm which makes the resulting embedding clusters less separated than in the exact version. Deviations from the optimal embedding could be magnified by the thresholding step, effectively making the stochastic algorithm more sensitive to the choice of threshold, and in effect the quality more sensitive to the choice of λ. In these experiments, the same threshold was used for both algorithms, but tailored choices could be considered given an appropriate selection criterion.

Furthermore, we see in Figure 2 that the running time of the stochastic splitting algorithm is lower than that of the exact algorithm, and grows significantly slower. While the stochastic splitting algorithm could in principle be implemented in time constant in the number of samples, and instead determined by the number of iterations, the adaptive stepsize used to improve performance requires evaluation of the objective value which scales with the number of samples. This could be improved by subsampling the terms in the objective function, but this was not done here.

6. Conclusions

We developed a stochastic incremental algorithm based on proximal iterations for the SON convex relaxtion of clustering that is highly suited to large scale problems and gave an analysis of its convergence propertis. We also gave quite general theoretical guaranteees for exact recovery of clusters similar to the unifying proximity condition in approximation algorithms that covers paradigm models for clustering data.

It has not escaped our attention that our algorithm can easily be adapted to incorporate similarity weights as used in Chi & Lange (2015); Chen et al. (2015); Hocking et al. (2011) and that it is amenable to acceleration using variance reduction and other techniques. The cluster recovery conditions can also be extended to cover almost perfect recovery i.e. correctly clustering all except a small fraction of points. A more complete experimental evaluation of our algorithm and comparison to others will be included in a longer version of the paper.

Acknowledgements

This work is supported in part by the Swedish Foundation for Strategic Research (SSF).

References

A. Kumar, R. Kannan. Clustering with spectral norm and the k-means algorithm. In *FOCS*, 2010.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. Optimization with sparsity-inducing penalties. *Foundation and Trends in Machine Learning*, 1(4):1–106, 2012.

Bertsekas, D. Incremental proximal methods for large scale convex optimization. *Math. Program.*, 129(163), 2011.

Bottou, Léon, Curtis, Frank E., and Nocedal, Jorge. Optimization methods for large-scale machine learning. Technical report, arXiv:1606.04838, June 2016. URL http://leon.bottou.org/papers/tr-optml-2016.

Chen, Gary K., Chi, Eric C., Ranola, John M.O., and Lange, Kenneth. Convex clustering: An attractive alternative to hierarchical clustering. *PLoS Computational Biology*, 11(5): e1004228, 2015. doi: 10.1371/journal.pcbi.1004228. URL http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004228.

Chi, Eric C. and Lange, Kenneth. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013, 2015. doi: 10.1080/10618600.2014.948181. URL http://dx.doi.org/10.1080/10618600.2014.948181.

Defazio, A., F. Bach, F., and Lacoste-Julien, S. A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pp. 1646–1654, 2014.

Hocking, T., Vert, J-P., Bach, F., and Joulin., A. Clusterpath: an algorithm for clustering using convex fusion penalties. In *ICML*, 2011.

Jain, A. K., Murty, M. N., and Flynn, P. J. Data clustering: A review. *ACM Comput. Surv.*, 31(3):264–323, September 1999. ISSN 0360-0300. doi: 10.1145/331499.331504. URL http://doi.acm.org/10.1145/331499.331504.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pp. 315–323, 2013.

Lindsten, F., Ohlsson, H., and Ljung, L. Clustering using sum-of-norms regularization: With application to particle filter output computation. 2011.

P.Awasthi, O Sheffet. Improved spectral norm bounds for clustering. In *APPROX-RANDOM*, 2012.

Rand, William M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

Schmidt, M., Roux, N. Le, and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 2016.

Zhu, C., Xu, H., Leng, C., and Yan, S. Convex optimization procedure for clustering: Theoretical revisit. In *NIPS*, 2010.