Learning Strictly Orthogonal $p$-Order Nonnegative Laplacian Embedding via Smoothed Iterative Reweighted Method

Haoxuan Yang$^{1*}$, Kai Liu$^{1*}$, Hua Wang$^1$ and Feiping Nie$^2$

$^1$Department of Computer Science, Colorado School of Mines, Golden, CO 80401, U.S.A.
$^2$School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi’an 710072, Shaanxi, P. R. China
haoxuan@mymail.mines.edu, {cskailiu, huawangcs, feipingnie}@gmail.com

Abstract

Laplacian Embedding (LE) is a powerful method to reveal the intrinsic geometry of high-dimensional data by using graphs. Imposing the orthogonal and nonnegative constraints onto the LE objective has proved to be effective to avoid degenerate and negative solutions, which, though, are challenging to achieve simultaneously because they are nonlinear and nonconvex. In addition, recent studies have shown that using the $p$-th order of the $\ell_2$-norm distances in LE can find the best solution for clustering and promote the robustness of the embedding model against outliers, although this makes the optimization objective nonsmooth and difficult to efficiently solve in general. In this work, we study LE that uses the $p$-th order of the $\ell_2$-norm distances and satisfies both orthogonal and nonnegative constraints. We introduce a novel smoothed iterative reweighted method to tackle this challenging optimization problem and rigorously analyze its convergence. We demonstrate the effectiveness and potential of our proposed method by extensive empirical studies on both synthetic and real data sets.

However, both positive and negative values in the solution of multi-way clustering tasks make the results hard to interpret directly, because the clustering indicator vectors require nonnegative results. In two-way clustering, this is not a problem, because a linear $\Psi$-transformation [Ding and He, 2004] of the eigenvectors leads to two genuine indicator vectors (each row has only one nonzero positive entry). Thus, mixed-sign solution is a generic difficulty for multi-way spectral clustering. To solve this problem, a clustering step usually has to be performed after the embedding is learned. That is, in traditional way, the clustering indicator vectors approximated by the eigenvectors of the Laplacian matrix will be grouped by using K-means clustering [Hartigan and Wong, 1979] in the eigenvector space. Thus, the traditional clustering solution provided from this process is neither stable nor intuitive, which may also be very sensitive to data outliers. To tackle this difficulty, Nonnegative Laplacian Embedding (NLE) method [Luo et al., 2009] was proposed by additionally imposing the nonnegative constraints on the embeddings.

Despite the fact that the nonnegativity can be achieved in the NLE method, there are still some difficulties of this model that are not well addressed. It has been noted that the NLE method imposes the nonnegative constraint at the price of relaxing the orthogonality on the learned approximations [Ding et al., 2006], although the orthogonality constraint ($X^TX = I$) is of significant importance to guarantee a good performance. The true meaning of the orthogonality constraint is to prevent degenerate solution ($X \rightarrow 0$). For one dimensional problem, the orthogonality can avoid that the embedded data collapse into a point. For multi-dimensional problem, the orthogonality can prevent data points from collapsing into a subspace with dimensions less than desired.

In this paper, we propose a new approach to learn LE with strictly guaranteed orthogonality and nonnegativity in the solution. Unlike using the auxiliary function method [Lee and Seung, 2001] to derive the solution algorithm for NLE in [Luo et al., 2009], the orthogonality of our solution is rigorously achieved by using the Alternating Direction Method of Multipliers (ADMM) [Bertsekas, 1996; Boyd et al., 2011], leading to a more stable solution and a better performance in the problem of spectral clustering. We also keep the nonnegativity in the constraint, such that the clustering membership can be readily read off from the embedded data due to the nonnegative constraint, i.e., we can
consider each row of the solution $X$ as the posterior clustering probability. In other words, the values in $i$-th row of the solution can be viewed as the likelihoods that the $i$-th data point belongs to different clusters, which gives our new approach the soft clustering capability that is crucial in many real-world applications.

Finally, we recognize that the squared $\ell_2$-norm distance used in the traditional LE and NLE objectives does not guarantee the optimal embedding [Wang et al., 2015] and is also notoriously known to be sensitive to the outliers [Wang et al., 2012; Nie et al., 2013; Wang et al., 2013c; Nie et al., 2016; Liu et al., 2017]. With strict orthogonality and nonnegative guaranteed simultaneously in the solution, we are also interested in promoting the robustness of our new NLE model by using the $p$-th order ($0 < p \leq 2$) of the $\ell_2$-norm distance in the objective. As a result, the proposed optimization objective is a quadratic function with both orthonormal and nonnegative constraints, which is highly nonlinear and nonconvex in its feasible domain. The $p$-th order term further makes the objective nonsmooth and difficult to efficiently optimize in general. To solve this challenging optimization problem, we propose a novel smoothed iterative reweighted method. Compared to the iterative reweighted method proposed in [Candes et al., 2008; Nie et al., 2010] to solve the $\ell_1$-norm or $\ell_2,1$-norm minimization problems, our new optimization framework explicitly adds a smoothness term which can improve numerical stability. Most importantly, as an important theoretical contribution, we rigorously prove the convergence of our new iterative algorithm with the smoothness term, which, though, was not present in [Candes et al., 2008; Nie et al., 2010] and their following works.

To evaluate the proposed robust NLE objective that uses the $p$-th order of the $\ell_2$-norm distances and our new smoothed iterative reweighted method, we have performed extensive empirical studies. The promising experimental results have validated the effectiveness of our new methods.

2 Strictly Orthogonal $p$-Order Nonnegative Laplacian Embedding

Given a set of $n$ data points, we can represent the pairwise similarities between these data points by a graph $G = \{V, E\}$, where the data points are represented by the vertices $V$ and $|V| = n$. Suppose that $W \in \mathbb{R}^{n \times n}$ denotes the affinity matrix of the graph $G$ where $w_{ij}$ measures the similarity between the $i$-th and the $j$-th vertices, quadratic placement [Hall, 1970] aims to embed the vertices of the graph into the one-dimensional space with coordinates $(x_1, \ldots, x_n)$, such that if the $i$-th and the $j$-th vertices are similar (i.e., $w_{ij}$ is large), they should be adjacent in embedded space, i.e., $(x_i - x_j)^2$ should be small. This can be achieved by the following objective [Hall, 1970]:

$$\min_{\|x\|_2^2 = 1} \sum_{i,j} w_{ij} (x_i - x_j)^2 = 2x^T(D - W)x \ ,$$

where $x = [x_1, \ldots, x_n]^T$, and $D = \text{diag}(d_1, d_2, \ldots, d_n)$ is the degree matrix of the graph with $d_i = \sum_j w_{ij}$.

The one-dimensional quadratic placement in Eq. (1) can be generalized to $r$-dimensional LE by minimizing the following objective [Luo et al., 2009]:

$$\min_{x \in \mathbb{C}} \sum_{i,j} w_{ij} \|x_i - x_j\|_2^2 = \text{tr} \left( X^T (D - W) X \right) \ ,$$

where $X = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^{n \times r}$. Obviously, the $i$-th row of $X$, i.e., $x_i^T \in \mathbb{R}^r$, is the embedding of the $i$-th data point in the $r$-dimensional space. Here, the orthonormal constraint of $X^T X = I$ is imposed in Eq. (2) to avoid degenerate solutions, which is critical as analyzed in [Luo et al., 2009; Ding et al., 2006].

To decode the clustering membership from $X$ in an easier way, Luo et al. [Luo et al., 2009] further developed LE by additionally imposing the nonnegative constraint onto the embedding matrix $X$ by minimizing the following objective:

$$\min_{X} \text{tr} \left( X^T (D - W) X \right) , \quad \text{s.t. } X \succeq 0, X^T X = I \ .$$

The squared $\ell_2$-norm distances used in the both objectives in Eq. (2) and Eq. (3) do not tolerate large value of distance, thus making the distances in the embedded space tend to be even, i.e., not too large but also not too small. Therefore, solving the objective in Eq. (2) or Eq. (3) may not find the optimal embedding such that most of the distances of local data pairs are minimized but a few of them are large [Wang et al., 2015]. Motivated by recent papers that use not-squared $\ell_2$-norm distances [Wang et al., 2012; Nie et al., 2013; Wang et al., 2013c; Wang et al., 2014; Nie et al., 2016; Liu et al., 2017] or the $p$-th order of the $\ell_2$-norm distances [Wang et al., 2011; Wang et al., 2013a; Wang et al., 2015] to promote the robustness of learning models, in this paper we propose to solve the following problem to find the optimal spectral embedding from an input graph:

$$\min_{X} \sum_{i,j} w_{ij} \|x_i - x_j\|_2^p , \quad \text{s.t. } X \succeq 0, X^T X = I \ ,$$

where $0 < p \leq 2$. Obviously, the NLE method in Eq. (3) [Luo et al., 2009] is a special case of our new method when $p = 2$. More importantly, by setting $p \leq 1$, the method will focus on minimizing most of the distances of local data pairs. Thus, we call Eq. (4) as the proposed strictly orthogonal $p$-Order Nonnegative Laplacian Embedding (PO-NLE) method.

3 Smoothed Iterative Reweighted Method and its Convergence

Although the motivation of our new objective in Eq. (4) is clear, it is nonsmooth and difficult to efficiently solve in general. To solve this challenging optimization problem, in this section we will first introduce a novel smoothed iterative reweighted method.

First, let us consider a general problem as follows:

$$\min_{x \in \mathbb{C}} f(x) + \sum_i \text{tr} \left( (g_i^T(x)g_i(x))^\frac{p}{2} \right) \ .$$

When $g_i(x)$ is a vector output function, $\text{tr} \left( (g_i^T(x)g_i(x))^\frac{p}{2} \right)$ becomes the following term:

$$\text{tr} \left( (g_i^T(x)g_i(x))^\frac{p}{2} \right) = \|g_i(x)\|_2^p \ .$$

4041
Equation (5) is nonsmooth, thus we turn to solve the following smooth problem [Liu et al., 2017]:
\[
\min_{x \in \mathcal{C}} f(x) + \sum_i \text{tr} \left( (g_i^T(x)g_i(x) + \delta I)^{\frac{p}{2}} \right),
\]
where \( \delta > 0 \) is a small constant. When \( \delta \to 0 \), Eq. (7) is reduced to Eq. (5) since the following equation holds:
\[
\lim_{\delta \to 0} \text{tr} \left( (g_i^T(x)g_i(x) + \delta I)^{\frac{p}{2}} \right) = \|g_i(x)\|^p_2.
\]

Before deriving the algorithm for optimizing the problem in Eq. (7), we need the following lemmas. First, according to the chain rule in calculus, we have:
\[
\frac{\partial h(g(x))}{\partial x} = \sum_i \frac{\partial h(g(x))}{\partial g_i(x)} \frac{\partial g_i(x)}{\partial x} = \left( \frac{\partial h(g(x))}{\partial g(x)} \right)^T \frac{\partial g(x)}{\partial x}.
\]

According to the chain rule in Lemma 1, we can easily derive the following two lemmas:

**Lemma 2** Suppose \( g(x) \) is a scalar output function, \( x \) is a scalar, vector or matrix variable, then we have:
\[
\frac{\partial \text{tr}(g^T(x)g(x) + \delta I)^{\frac{p}{2}})}{\partial x} = p(g^T(x)g(x) + \delta I)^{\frac{p-2}{2}} g^T(x) \frac{\partial g(x)}{\partial x}.
\]

**Lemma 3** Suppose \( g(x) \) is a scalar, vector or matrix output function, \( x \) is a scalar, vector or matrix variable, \( D \) is a constant and \( D \) is symmetrical if \( D \) is a matrix, then we have:
\[
\frac{\partial \text{tr}(g^T(x)g(x)D)}{\partial x} = 2Dg^T(x) \frac{\partial g(x)}{\partial x}.
\]

Now we derive the algorithm to optimize the problem in Eq. (7). The Lagrangian function of the problem in Eq. (7) is:
\[
\mathcal{L}(x, \lambda) = f(x) + \sum_i \text{tr}((g_i^T(x)g_i(x) + \delta I)^{\frac{p}{2}}) - \lambda r(x, \lambda),
\]
where \( r(x, \lambda) \) is a Lagrangian term for the constraint \( x \in \mathcal{C} \). By setting the derivative of Eq.(12) w.r.t. \( x \) to zero, we have
\[
\frac{\partial \mathcal{L}(x, \lambda)}{\partial x} = f'(x) + \sum_i \text{tr}((g_i^T(x)g_i(x) + \delta I)^{\frac{p}{2}}) - \frac{\partial r(x, \lambda)}{\partial x} = 0.
\]

According to Lemma 2, Eq.(13) can be rewritten as
\[
f'(x) + \sum_i p(g_i^T(x)g_i(x) + \delta I)^{\frac{p-2}{2}} g_i^T(x) \frac{\partial g_i(x)}{\partial x} - \frac{\partial r(x, \lambda)}{\partial x} = 0.
\]

**Algorithm 1** The algorithm to solve the problem (7)

Initialize \( x \in \mathcal{C} \).

while not converge do

1. For each \( i \), calculate
\[
D_i = \frac{p}{2}(g_i^T(x)g_i(x) + \delta I)^{\frac{p-2}{2}},
\]

2. Update \( x \) by solving the problem:
\[
\min_{x \in \mathcal{C}} f(x) + \sum_i \text{tr}(g_i^T(x)g_i(x)D_i).
\]
end while

If we can find a solution \( x \) that satisfies the Eq.(14), we usually find a local or global optimal solution to the problem in Eq. (7) according to the Karush-Kuhn-Tucker (KKT) conditions [Boyd and Vandenberghe, 2004]. However, directly finding a solution \( x \) that satisfies Eq.(14) is not an easy task. In this paper, following [Candes et al., 2008; Nie et al., 2010] we propose an iterative algorithm to find it using the following observation: if
\[
D_i = \frac{p}{2}(g_i^T(x)g_i(x) + \delta I)^{\frac{p-2}{2}}
\]
is a constant, Eq.(14) is reduced to:
\[
f'(x) + \sum_i 2D_i g_i^T(x) \frac{\partial g_i(x)}{\partial x} - \frac{\partial r(x, \lambda)}{\partial x} = 0,
\]
which is equivalent to solving the following problem:
\[
\min_{x \in \mathcal{C}} f(x) + \sum_i \text{tr}(g_i^T(x)g_i(x)D_i).
\]

Based on the observation above, we can first guess a solution \( x \). Then we calculate \( D_i \) using the current solution of \( x \) and update \( x \) by the optimal solution of the problem in Eq. (17) by calculating \( D_i \). We iteratively perform this procedure until it converges, which is summarized in Algorithm 1.

The convergence of Algorithm 1 is guaranteed by the following theorem. The proof of Theorem 1 will be supplied in the extended journal version of this paper due to space limit.

**Theorem 1** The Algorithm 1 will monotonically decrease the objective of the problem (7) in each iteration until the algorithm converges.

In the convergence, the equality in Eq. (14) will hold, thus the KKT condition of problem (7) is satisfied. Therefore, the Algorithm 1 will converge to a local optimum solution to the problem (7). If the problem (7) is convex, the Algorithm 1 will converge to a global optimum solution.

Here we note that the iterative reweighted method introduced in [Candes et al., 2008; Nie et al., 2010] solves the nonsmooth \( \ell_1 \)-norm or \( \ell_{2,1} \)-norm minimization problems. However, the method described in [Candes et al., 2008; Nie et al., 2010] does not explicitly use the smoothness constant (i.e., \( \delta I \) in Eq. (7)). Without this smoothness term,
the algorithm is heavily impacted by the singularity problem due to inverted matrices that divide 0s, which routinely leads to inferior learning performances. To improve the numerical stability, in [Nie et al., 2010] and the following works by the same group of authors [Wang et al., 2013b; Nie et al., 2013], a smoothness term was informally added for empirical purpose. But they only theoretically proved the convergence of the algorithm that does not use the smoothness term and did not provide any theoretical analysis on the objectives that use the smoothness term. As an important theoretical contribution of this paper, we formally introduce the smoothness term (i.e., \( S \) in Eq. (7)) into our algorithm and theoretically prove the convergence of our algorithm in which the smoothness term leads to much more stable solutions. We call Algorithm 1 as the proposed Smoothed Iterative Reweighted Method, which can be broadly used to solve a variety of difficult machine learning problems that minimize the objectives using the \( p \)-th order of \( \ell_2 \)-norm distances, the \( p \)-th order of \( \ell_p \)-norm distances, or the \( p \)-th order of the Schatten \( p \)-norm.

4 Algorithm to Solve the Problem in Eq. (4)

Equipped with Algorithm 1, we can derive the solution algorithm to our new PO-NLE objective in Eq. (4) now. According to Step 2 of Algorithm 1, the key step to solve Eq. (4) is to solve the following problem:

\[
\min_X \sum_{i,j} w_{ij} d_{ij} \| x_i - x_j \|^2_2, \quad s.t. \ X \succeq 0, \ X^T X = I, \tag{20}
\]

where \( d_{ij} = \frac{1}{2} \left( \| x_i - x_j \|^2_2 + \delta \right)^{\frac{p}{2}} \) and \( \delta \to 0 \).

Denote \( \tilde{w}_{ij} = w_{ij} d_{ij} \) and let \( \tilde{D} \) be the diagonal matrix with the \( i \)-th diagonal entry as \( \sum_j \tilde{w}_{ij} = w_{ij} d_{ij} \). The problem in Eq. (20) can be written as following:

\[
\min_X \sum_{i,j} \tilde{w}_{ij} \| x_i - x_j \|^2_2 = \text{tr} \left( X^T \left( \tilde{D} - \tilde{W} \right) X \right), \quad \tag{21}
\]

\[\quad \text{s.t. } \ X \succeq 0, \ X^T X = I.\]

Obviously, Eq. (21) is identical to the NLE objective in Eq. (3), which was proposed in [Luo et al., 2009]. In [Luo et al., 2009], a solution algorithm was derived using the auxiliary function method [Lee and Seung, 2001]. However, as analyzed in [Ding et al., 2005; Ding et al., 2006] the orthogonal constraint indeed are not guaranteed, which, though, is very important to avoid degenerate solutions [Ding et al., 2006]. Thus, instead of using the solution algorithm provided in [Luo et al., 2009], we derive the solution algorithm to solve Eq. (21) using the ADMM method.

Denoting \( L = \tilde{D} - \tilde{W} \) for brevity\(^1\), we can write the objective in Eq. (21) as following:

\[
\min_X \text{tr} \left( X^T LX \right), \quad \text{s.t. } X^T X = I, X \succeq 0 \quad \tag{22}
\]

We can solve Eq. (22) by solving the following equivalent optimization problem:

\[
\min_{X,Y} \text{tr} \left( Y^T LX \right), \quad \text{s.t. } Y = X, Y^T Y = I, X \succeq 0, \tag{23}
\]

where the constraint of \( X^T X = I \) in Eq. (22) is implicitly enforced by the constraints of \( Y = X \) and \( Y^T Y = I \).

Following the ADMM optimization framework, we need to solve the following optimization problem:

\[
\min_{X,Y,A} \text{tr} \left( Y^T LX \right) + \frac{\mu}{2} \left\| Y - X + \frac{1}{\mu} A \right\|_F^2, \quad \tag{24}
\]

\[\text{s.t. } Y^T Y = I, X \succeq 0, \]

in which we introduced the Lagrangian multiplier \( A \) for the constraint of \( Y = X \). The detailed procedures to solve Eq. (24) using the ADMM method will be supplied in the extended journal version of this paper due to space limit.

5 Experiment and Results

In this section we empirically evaluate our new PO-NLE method on one synthetic data set, four data sets from the UCI Machine Learning Data Repository, and three image data sets. We will compare our new method against its counterparts: NLE, Normalized Cut (NCut) [Shi and Malik, 2000] and Laplacian Embedding (LE).

In our evaluations, we use clustering accuracy and clustering purity to measure the performance of the compared methods. We also study the robustness of our method on the real-world data sets when they are contaminated with noise. The performance variations when we increase the value of \( p \) will be shown to validate our hypothesis that the optimal solution is usually obtained when \( p \) is less than 2 and close to 1 (it depends on data sets), given that noises are present in the data. Orthogonality of the solution will be illustrated and compared against the NLE method in [Luo et al., 2009].

5.1 Experiments on a Synthetic Data Set

To illustrate the effectiveness of our new PO-NLE method, we create a synthetic data set as follows. We first randomly generate 3 data points as centroids in the 30-dimensional space. Then we generate 3 groups of data points and each group consists of 39 data points which are randomly distributed around one of the three centroids. A threshold is set to make the distance of groups large enough. As shown in Figure 1, different colors (red, black and blue) and shapes are used to represent different groups of data points. We randomly initialize \( X (0 \leq X \leq 1) \) and set \( \rho = 1.02, \mu = 0.1 \) and \( p = 0.8 \) in our algorithm. We use K-Nearest Neighbors with heat kernel to construct our adjacency matrix \( W \). The variation of the objective value when our algorithm iterates are shown as the red curve in Figure 1. For visualization purpose, we set \( r = 3 \), i.e., we embed the original data into the 3-dimensional space using our new PO-NLE algorithm. The \( x, y \) and \( z \) axes of the 3D plots in the figure correspond to the first, second and third row in matrix \( X \), respectively.

From Figure 1, we observe that the objective function monotonically decreases in each iteration, which empirically

---

\(^1\)In practice, due to the zero mode of the Laplacian matrix of a graph [Wang et al., 2010], we compute \( L = D - W + \frac{w_{++}}{\rho^2} e^T e \) to ensure that \( L \) is positive definite, where \( W_{++} = \sum_{i,j} W \) and \( e \) is the vector with all entries to be 1.
confirms the convergence of the solution algorithm to solve PO-NLE derived by our new smoothed iterative reweighted method. Moreover, for each checkpoint shown by the black circles on the objective curve, the clustering structure of the experimental data becomes more and more clear in the 3D plots when the algorithm iterates. The three clusters of data points gradually find a solution to separate themselves apart and fall on different axes. Note that, due to the nonnegative constraints on $X$, data points will finally converge on the positive part of each axis. This observation clearly demonstrate the effectiveness of the proposed new method.

5.2 Studies of the Orthogonality of the Solutions of Our New Method

An important improvement of our new method over the NLE method [Luo et al., 2009] is that the orthogonality of our solution is rigorously guaranteed, which, as analyzed in [Ding et al., 2005; Ding et al., 2006] is very important to avoid degenerate solutions. Thus, in this subsection we empirically study the orthogonality of the solutions of our new method and compare them against the solutions obtained from the NLE method. Figure 2 compare the visualizations of $X^T X$ learned from the two compared methods on the Glass data set by the heatmaps. The heatmap of our new method is on the top and that of the NLE method is at the bottom. From the results we can see that the learned embeddings from our method are strictly orthogonal as shown in Figure 2(a), which will in return lead to better clustering performances and robustness after embedding. In contrast, the NLE method failed to guarantee the orthogonality, as can be seen in Figure 2(b).

5.3 Experiments on Noiseless Real Data Sets

Now we compare our new method, NLE, NCut and LE on the seven standard data sets as summarized in Table 1. Each data set will be tested by different algorithms independently for 200 times. For NCut and LE algorithms, we run K-means clustering with random initialization for 50 times and report the best results.

The performances of the compared methods are reported in the top half of Table 2, from which we can see that our method clearly outperforms all other competing methods, especially on those comparatively noisier data sets. Due to the nonnegative solutions of our new method, we do not need any additional clustering step. Instead, the clustering membership can be readily read off directly from the learned embeddings. The strictly guaranteed orthogonality constraint avoids degenerate solution and helps improve the performance com-

Table 1: Dataset descriptions.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Size</th>
<th># Dimension</th>
<th># Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>MINIST</td>
<td>5000</td>
<td>784</td>
<td>10</td>
</tr>
<tr>
<td>AT&amp;T</td>
<td>400</td>
<td>10304</td>
<td>40</td>
</tr>
<tr>
<td>Caltech101</td>
<td>332</td>
<td>900</td>
<td>5</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>351</td>
<td>34</td>
<td>2</td>
</tr>
<tr>
<td>Wine</td>
<td>178</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Iris</td>
<td>150</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Glass</td>
<td>214</td>
<td>9</td>
<td>6</td>
</tr>
</tbody>
</table>
pared with loosely constrained NLE method which does not have such desirable property. To illustrate the convergence of the objective function of our new method, Figure 3(a) and Figure 3(b) show a typical run of our algorithm on two UCI benchmark data sets. As can be seen from the figures, when the algorithm iterates and the objective value decreases, the accuracy shows a relatively smoothly increasing line.

5.4 Experiments on Noisy Real Data Sets

To study the impacts of the value of $p$ in our new embedding model, we randomly contaminate 20% of the data points in all 7 data sets and we run our method with increasing $p$ on those data sets. For each $p$, we run 200 times for the same contaminated data and original data respectively. Other algorithms are also tested for 200 times on each data set for comparison. The performances of the clustering methods on contaminated noisy data sets are reported in the bottom half of Table 2. Among all the best and the average values of clustering accuracy and clustering purity, our method is consistently better than its counterparts. The results of our approach generally decreases less than other methods on the contaminated data.

6 Conclusion

In this paper, we proposed a new robust Laplacian embedding approach that uses the $p$-th order of the $\ell_2$-norm distances in the objective and strictly satisfies both orthogonality and nonnegativity constraints at the same time. This results in an objective that is neither convex nor smooth, which is difficult to efficiently solve in general. We thereby proposed a novel smooth iterative reweighted method to solve this challenging optimization problem, in which a smoothness term is formally and explicitly introduced for improved numerical stability. As an important theoretical contribution of this paper, we rigorously proved the convergence of our new algorithm with the smoothness term. Using this new and improved optimization framework, our objective can be elegantly solved. We have performed extensive experiments, in which the superior performance of our new method has demonstrated its effectiveness and the potential to give a new perspective for nonlinear graph based clustering tasks.

Acknowledgments

All correspondence should be addressed to: Hua Wang (huawangcs@gmail.com). This work was partially supported by National Science Foundation under Grants IIS-1652943 and IIS-1849359.

Table 2: Best and average (Ave) clustering accuracy and purity by our method, NLE, NCut and LE over 200 trials. "\top" means that the bigger number are the better. 

![Figure 3: A typical run of our algorithm on two data sets with iteration ranging from 1 to 300 to illustrate the convergence of objective function and accuracy of the clustering result.](image-url)
References


