Beyond the Simplex: Hadamard-Infused Deep Sparse Representations for Enhanced Similarity Measures

Xiangyu Li, Umberto Gherardi
Department of Computer Science
Colorado School of Mines
Golden, Colorado, USA
{lixiangyu, umbertogherardi}@mines.edu

Armand Ovanessians
Department of Quantitative Biosciences and Engineering
Colorado School of Mines
Golden, Colorado, USA
ovanessians@mines.edu

Hua Wang
Department of Computer Science
Colorado School of Mines
Golden, Colorado, USA
huawangcs@gmail.com

2023 IEEE International Conference on Knowledge Graph (ICKG) | 979-8-3503-0709-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICKG59574.2023.00026

***Abstract*—Graphical representations are essential for comprehending high-dimensional data across diverse fields, yet their construction often presents challenges due to the limitations of traditional methods. This paper introduces a novel methodology, Beyond Simplex Sparse Representation (BSSR), which addresses critical issues such as parameter dependencies, scale inconsistencies, and biased data interpretation in constructing similarity graphs. BSSR leverages the robustness of sparse representation to noise and outliers, while incorporating deep learning techniques to enhance scalability and accuracy. Furthermore, we tackle the optimization of the standard simplex, a pervasive problem, by introducing a transformative approach that converts the constraint into a smooth manifold using the Hadamard parametrization. Our proposed Tangent Perturbed Riemannian Gradient Descent (T-PRGD) algorithm provides an efficient and scalable solution for optimization problems with standard simplex or ℓ_1-norm sphere constraints. These contributions, including the BSSR methodology, robustness and scalability through deep representation, shift-invariant sparse representation, and optimization on the unit sphere, represent major advancements in the field. Our work offers novel perspectives on data representation challenges and sets the stage for more accurate analysis in the era of big data.**

***Index Terms*—Data Similarity, Sparse Representation, Simplex Constraint, Riemannian Optimization.**

I. Introduction

Graphical representations play a crucial role in understanding high-dimensional data across various fields, including scientific computing, machine learning, and information technology. However, the construction of these graphical structures often presents challenges due to the limitations of traditional methods. These methods struggle with tasks such as selecting suitable thresholds for the ϵ-neighbor network, determining the optimal number of neighbors for the k-nearest nodes graph, and defining appropriate similarity functions for the fully connected graph. These challenges arise from parameter dependencies, scale inconsistencies, and the symmetric treatment of similarity matrices, potentially leading to biased data interpretation.

In the realm of similarity graph construction, many applications rely on finding suitable approaches that address specific requirements based on the dataset and application. However, several challenges in this domain remain unresolved. These include determining the optimal scale of analysis, selecting the appropriate number of neighbors, handling multi-scale data, and effectively managing noise and outliers. Although notable advancements have been made in addressing some of these challenges, such as [4]–[6], no single method currently comprehensively tackles all of these challenges to the best of our knowledge. Therefore, further research is needed to develop holistic solutions that encompass all aspects of similarity graph construction.

To address these challenges, we propose a novel methodology called Beyond Simplex Sparse Representation (BSSR) for constructing similarity graphs. Our approach leverages the robustness of sparse representation to noise and outliers, without imposing restrictions on the scale consistency of data vectors. Building upon the sparse representation framework proposed in [10], we compute the similarity matrix S. The BSSR method is robust, parameter-independent, and takes into account the possibility of asymmetrical relationships in the similarity matrix. By harnessing the power of sparse representation and deep learning techniques, we effectively handle high-dimensional data, resulting in reliable and interpretable graphical representations, and facilitating efficient downstream clustering processes.

Furthermore, in our proposed objective, we encounter the challenge of optimizing the standard simplex, a prevalent problem across various fields. Traditional Projected Gradient Descent (PGD) struggles with this problem due to the non-

979-8-3503-0709-2/23/$31.00 ©2023 IEEE
DOI 10.1109/ICKG59574.2023.00026

smooth nature of the simplex constraints. To overcome this limitation, we introduce a transformative approach that converts the standard simplex constraint into residing on the unit sphere using the Hadamard parametrization [11], [18]. This conversion effectively transforms the constrained optimization problem into a smooth and simple manifold.

Through rigorous theoretical analysis, we establish a profound connection between the original problem and the transformed problem. We demonstrate that the KKT points and strict-saddle points of the original problem correspond to those of the transformed problem, ensuring their mutual solvability. Building upon this transformative framework, we propose an efficient algorithm called Tangent Perturbed Riemannian Gradient Descent (T-PRGD), which leverages the manifold structure to address the optimization problem. The T-PRGD algorithm provides an effective and scalable solution for optimization problems with standard simplex or ℓ_1-norm sphere constraints.

We believe that our work contributes significantly to the field, offering fresh perspectives on the challenges of data representation and paving the way for more accurate analysis in the era of big data. Our contributions encompass:

1) Beyond Simplex Sparse Representation (BSSR): Our parameter-independent approach revolutionizes data analysis by introducing a reliable and simplified technique that surpasses traditional methods.
2) Robustness and Scalability through Deep Learning: By integrating deep learning, we enhance the robustness of our method to scale inconsistencies and outlier noise, enabling scalability for complex datasets.
3) Shift-Invariant Sparse Representation: Our method incorporates a simplex constraint into sparse representation, ensuring shift-invariance and promoting sparser representations. This enhances data interpretation accuracy and computational efficiency.
4) Optimization on the Unit Sphere: We propose an innovative reparametrization method that optimizes the standard simplex problem by transforming it onto the Riemannian manifold of the unit sphere. Our Tangent Perturbed Riemannian Gradient Descent (T-PRGD) technique improves efficiency, robustness, and accuracy, demonstrating our commitment to pioneering optimization solutions.

II. Formulation and Algorithm

Suppose we have m data vectors of size d, arranged as columns in a training sample matrix $X = (x_1, \ldots, x_m) \in \mathbb{R}^{d\times m}$. The objective is to obtain a sparse and non-negative representation for each data point with respect to the remaining points. This task is commonly addressed through sparse coding or sparse representation algorithms [10], [13], [15], which enable the calculation of pairwise similarities between the data points:

$$\min_{s_i\geq 0} \sum_{i=1}^{m} \left(||X_{-i}s_i - x_i||_2^2 + \lambda ||s_i||_1 \right). \tag{1}$$

In Eq (1), $X_{-i} \in \mathbb{R}^{d\times(m-1)}$ denotes the data matrix excluding the i-th column, effectively representing all other data points. The vector $s_i \in \mathbb{R}^{m-1}$, subject to a non-negativity constraint based on the assumption that the similarity matrix is usually non-negative, is the sparse representation coefficient for the i-th data point. It describes the linear combination of other data points that approximates the i-th data point $x_i \in \mathbb{R}^d$.

Addressing potential asymmetry in our similarity matrix $S = [s_1, \ldots, s_m]$, where s_i signifies the similarity coefficient assigned to the i-th data point, we introduce a symmetry-inducing operation. We rectify the matrix by averaging S and its transpose to obtain a symmetric similarity matrix, computed as: $W = \frac{(S+S^T)}{2}$. This step mitigates discrepancies between the similarity coefficients s_{ij} and s_{ji}, ensuring more accurate data representation. With the symmetric matrix W, we can confidently proceed with conventional clustering procedures such as Laplacian matrix computation and k-means clustering.

The objective function in Equation (1) comprises two components. The first term, $||X_{-i}s_i - x_i||_2^2$, measures the reconstruction error, which quantifies the Euclidean distance between the original data point x_i and its approximation using the other data points. The regularization term, $\lambda||s_i||_1$, uses the ℓ_1-norm regularization that promotes sparsity [8], [12], [14] in the representation by encouraging solutions with fewer non-zero components in s_i. This term introduces a trade-off between sparsity and reconstruction error. A higher value of λ increases the emphasis on sparsity, potentially leading to a sparser representation but higher reconstruction error. Conversely, a smaller value of λ prioritizes minimizing the reconstruction error, potentially resulting in a less sparse representation.

However, this approach has been impeded by two significant limitations. The first is the inherent assumption of a linear relationship among data points, which restricts the ability of this approach to capture complex, non-linear relationships inherent in many data structures. The second limitation pertains to scalability. The computational complexity of the traditional approach, which involves the minimization of a sum of functions for each data point, escalates rapidly with an increase in data volume, making it impractical for handling large datasets.

To counter these limitations, we introduce a new learning-based objective function that extends the original sparse representation paradigm by incorporating a non-linear transformation. This transformation, learned by a deep neural network, facilitates the exploration of richer, high-dimensional representations of the data, thus enhancing our ability to depict intricate, non-linear correlations within the data.

Our proposed learning-based objective function is formulated as:

$$\min_{s_i\geq 0,\theta} \sum_{i=1}^{m} \left(||\Theta(X_{-i};\theta)s_i - x_i||_2^2 + \lambda ||s_i||_1 \right). \tag{2}$$

In this formulation, $\Theta(X_{-i};\theta)$ denotes a deep neural network transformation of the data matrix X_{-i}, excluding the i-th column. This allows for nonlinear exploration of the data,

capturing complex relationships between data points. The term $\Theta(X_{-i};\theta)s_i$ represents the approximation of the i-th data point x_i using the transformed representations of the other data points. The integration of deep learning techniques not only enhances representation capabilities but also improves scalability. By utilizing stochastic optimization methods, like mini-batch gradient descent, the approach achieves computational efficiency and scalability. This is particularly advantageous for large-scale datasets.

In the original sparse representation-based approach, the computed similarities are also sensitive to constant shifts in the data, which could lead to potential inaccuracies or inconsistencies in data interpretation. When the data points are shifted by a constant vector $t = [t_1, \ldots, t_m]^T$, such that $x_k = x_k + t$ for any k, the similarities change accordingly. Ensuring shift-invariance, therefore, is crucial for maintaining the reliability of our analyses. To obtain shift-invariant similarities, the following equation needs to be satisfied:

$$||(X_{-i} + t1^T)s_i - (x_i + t)||_2^2 = ||X_{-i}s_i - x_i||_2^2. \quad (3)$$

This equation indicates that the sum of the coefficients in the sparse representation for each data point, $s_i^T 1$, is equal to 1. By incorporating this constraint into our earlier objective function, we reformulate the optimization problem as follows:

$$\min_{s_i,\theta} \sum_{i=1}^{m} \left(||\Theta(X_{-i};\theta)s_i - x_i||_2^2 + \lambda||s_i||_1\right) \text{ s.t. } s_i \geq 0, s_i^T 1 = 1 \quad (4)$$

The constraints in this optimization problem enforce a simplex structure on the sparse representation, ensuring that the sum of elements in the vector s_i is equal to 1. This structure encourages sparsity by allowing a few non-zero elements in s_i to have larger values, while maintaining the constraint. Consequently, the ℓ_1-norm regularization term, initially included to induce sparsity, becomes unnecessary in the presence of the simplex constraint. This approach, known as the "simplex representation", utilizes the unique properties of the simplex structure to promote sparsity.

To integrate the concept of simplex representation into our learning-based objective function, we revise the objective as follows:

$$\min_{s_i,\theta} \sum_{i=1}^{m} \left(||\Theta(X_{-i};\theta)s_i - x_i||_2^2\right) \quad \text{s.t. } s_i \geq 0, s_i^T 1 = 1. \quad (5)$$

This formulation combines the shift-invariance property with the power of learning-based models, enabling us to capture intricate data relationships while preserving the desirable properties of the similarity matrix.

However, optimizing within the simplex constraints presents challenges. Traditional approaches, such as Projected Gradient Descent (PGD), can be computationally intensive for complex objective functions. To overcome this, we propose a transformative methodology that reshapes the optimization landscape, leading to more efficient optimization.

We achieve this transformation by reparametrizing our vector s_i via Hadamard (element-wise) multiplication as $s_i = z_i \circ z_i$, where z_i resides on the unit sphere, designated as S^{n-1}, where $S^{n-1} := \{z \in \mathbb{R} : ||z||_2 = 1\}$ is the unit sphere. This leads to the transformed objective function:

$$\min_{\theta, z_i \in S^{n-1}} \sum_{i=1}^{m} \left(||\Theta(X_{-i};\theta)(z_i \circ z_i) - x_i||_2^2\right). \quad (6)$$

By transforming the simplex-constrained problem to reside on the unit sphere, we significantly simplify the optimization task. This transformation brings the problem onto a smooth manifold, a space without edges or discontinuities, which offers substantial benefits for optimization. Within this space, we can more easily calculate derivatives, which in turn allows for more efficient and robust computation of optimization algorithms. The unit sphere, with its smooth, continuous surface, serves as an excellent domain for our transformed optimization problem.

III. Algorithm and Theoretical Analysis

A. Riemannian Optimization

As we optimize over the set of points residing on the unit sphere, s_i^{n-1}, this constraint set forms a Riemannian manifold, a smooth manifold with an inner product that varies smoothly from point to point. This calls for the application of Riemannian optimization techniques that adjusts traditional optimization methods such as gradient descent or second-order methods to the geometry of the manifold. To address this requirement, we introduce an innovative Riemannian optimization method, termed Tangent Perturbed Riemannian Gradient (T-PRGD), devised to solve our problem on the unit sphere, s_i^{n-1}, as shown in Algorithm 1.

Algorithm 1 T-PRGD

Input: s: initial point, α: learning rate, β: perturbation scale, K: number of iterations, g: transformed objective function, Θ: neural network, θ: parameters of the neural network, X, x_i: data.
$z_0 = \sqrt{s}$
$g(z,\theta) := \sum_{i=1}^{m} \left(||\Theta(X_{-i};\theta)(z \circ z) - x_i||_2^2\right)$
Initialize the neural network parameters θ
for $k = 1, ..., K$ **do**
    Update parameters θ by minimizing $g(z_k, \theta)$
    Compute the gradient $\nabla g(z_k, \theta)$ with respect to z_k
    $\epsilon_k =$ random perturbation with scale β
    $v_k = \nabla g(z_k, \theta) + \epsilon_k$ (Update with perturbation)
    $z_{k+1} = \exp_{s_k}(-\alpha v_k)$ (Update z_k)
end for
Return $s_K = z_K \circ z_K$, θ

For the i-th data point, we denote the reparametrized vector as $z_i \in s_i^{n-1}$. At each point z_i, we define a corresponding tangent space $T_{z_i}s_i^{n-1} = \{v \in \mathbb{R}^n : v^\top z_i = 0\}$, consisting of all vectors orthogonal to z_i. We introduce a projection operator $Proj_{z_i}$ to project vectors from the ambient space onto the tangent space $T_{z_i}s_i^{n-1}$. For a given vector $w \in \mathbb{R}^n$, the projection operation becomes $Proj_{z_i}(w) = w - (w^\top z_i)z_i$.

Our problem's transformed objective function in Eq.(6) is defined as $\boldsymbol{g(z_i)}$. At any point $z_i \in s_i^{n-1}$, we designate the Riemannian gradient as the projection of the Euclidean

gradient onto the tangent space $T_{z_i}s_i^{n-1}$, denoted as grad$z_i g = Proj_{z_i}\nabla g(z_i)$. The Riemannian Hessian at a point $z_i \in s_i^{n-1}$ is formulated as the operator Hess$g(z_i) = Proj_{z_i} \circ (\nabla^2 g(z_i) - \nabla g(z_i)^\top z_i) \circ Proj_{z_i}$.

With a given $z_i \in s_i^{n-1}$ and a tangent vector $v \in T_{z_i}s_i^{n-1}$, we define the geodesic mapping at z_i in the direction v as $\gamma_{z_i,v}(t) : \mathbb{R} \to s_i^{n-1}$. Additionally, the exponential map at z_i translates a tangent vector to a point on the sphere along the geodesic direction, represented as exp$z_i : T_{z_i}s_i^{n-1} \to s_i^{n-1}$, mapping $v \mapsto \gamma_{z_i,\hat{v}}(|v|)$, where $\hat{v} = v/|v|$.

To iteratively minimize the transformed objective function, we employ the Tangent Riemannian Gradient Descent (T-RGD). Unlike Euclidean gradient descent, T-RGD utilizes the Riemannian gradient and traverses along geodesics instead of straight lines. Denoting η_k as the step size at the k-th iteration, the T-RGD update rule in our context becomes:

$$z_{i,k+1} = \exp_{z_{i,k}}(-\eta_k \text{grad } g(z_{i,k})). \tag{7}$$

This update ensures that the reparametrized vectors $z_{i,k+1}$ consistently reside on the sphere s_i^{n-1}, verifying the feasibility of T-RGD for our problem.

It's worth noting that the transformed objective function $\boldsymbol{g(z_i)}$ inherits the smoothness from the original function, as given by Eq.(5). For the ease of our discussion, we refer to the original problem by using the notation $\boldsymbol{f(s_i)}$. Specifically, if $f(s_i)$ exhibits L-Lipschitz differentiability, the corresponding characteristic of $g(z_i)$ is denoted as $\tilde{L}$-Lipschitz differentiability, where $\tilde{L} = 4L+2M$. Here, M stands for the supremum of the infinity norm of the gradient of $f(s_i)$, calculated over all s in the set Δ_n, that is, $M = \sup_{s\in\Delta_n} \|\nabla f(s_i)\|_\infty$, where Δ_n denotes the set $\{s_i \in \mathbb{R}^{m-1} : s_i \geq 0, \text{and } \mathbf{1}^T s_i = 1\}$. Given the continuity of $\nabla f(s_i)$ and the compactness of the domain Δ_n, it follows that $M < \infty$. The specifics of this characteristic are elaborated in Lemma 1. This finding ensures the requisite attributes of $g(z_i)$ for the successful application of RGD.

Lemma 1. *If f is L-Lipschitz differentiable, then the transformed objective function g is Lipschitz differentiable with Lipschitz constant $\tilde{L} = 4L + 2M$, where M is the supremum of the ℓ_2-norm of the gradient of f, over all x in Δ_n, i.e., $M = \sup_{x\in\Delta_n} \|\nabla f(s)\|_2$.*

Proof. We begin by considering the gradient of the function g, with respect to z, given by $\nabla_z g(z) = 2\nabla_s f(z \circ z) \circ z$. By exploring the difference between the gradients of g at two distinct points, z_1 and z_2, we arrive at the following inequality:

$$\|\nabla g(z_1) - \nabla g(z_2)\|_2 = 2\|\nabla_s f(z_1 \circ z_1) \circ z_1 - \nabla_s f(z_2 \circ z_2) \circ z_2\|_2.$$

Proceeding, we bound this difference by applying the triangle inequality and the Lipschitz condition for $\nabla f(s)$:

$$\|\nabla g(z_1) - \nabla g(z_2)\|_2 \leq 2\|\nabla_s f(z_2 \circ z_2) \circ (z_1 - z_2)\|_2 + 2\|\nabla_s f(z_1 \circ z_1) \circ z_1 - \nabla_s f(z_2 \circ z_2) \circ z_1\|_2.$$

The first term is constrained by L (the Lipschitz constant), and the second term by the supremum M, which will be justified in the next following lemma.

$$\|\nabla g(z_1) - \nabla g(z_2)\|_2 \leq 2L\|2(z_1 \circ z_1 - z_2 \circ z_2)\|_2 + 2M\|z_1 - z_2\|_2.$$

In the end, we establish that the inequality is less or equal to $\hat{L}\|z_1 - z_2\|_2$, with $\hat{L} = 4L + 2M$:

$$\|\nabla g(z_1) - \nabla g(z_2)\|_2 \leq \hat{L}\|z_1 - z_2\|_2$$

This derivation demonstrates that the Lipschitz constant of the gradient of the transformed function g is indeed $\hat{L}$, as stated in this lemma. □

Lemma 2. *Suppose we have two points z_1 and z_2 in the $(n-1)$-dimensional unit sphere s_i^{n-1}. We can measure the difference between these points in terms of their element-wise squared values. Specifically, the ℓ_2-norm of the difference between $z_1 \circ z_1$ and $z_2 \circ z_2$ does not exceed twice the ℓ_2-norm of the difference between z_1 and z_2. We can mathematically represent this relationship as $\|z_1 \circ z_1 - z_2 \circ z_2\|_2 \leq 2\|z_1 - z_2\|_2$.*

Proof.

$$\|z_1 \circ z_1 - z_2 \circ z_2\|_2 \leq \|z_1 \circ (z_1 - z_2)\|_2 + \|(z_1 - z_2) \circ z_2\|_2 \leq \|z_1\|_\infty \|z_1 - z_2\|_2 + \|z_1 - z_2\|_2 \|z_2\|_\infty \leq 2\|z_1 - z_2\|_2.$$

This sequence of inequalities is based on the key property that for any vectors a and b, the ℓ_2-norm of their Hadamard product can be bounded by the product of the ℓ_2-norm of one vector and the infinity norm of the other, i.e., $\|a \circ b\|_2 \leq \|a\|_2 \|b\|_\infty$. □

By incorporating these Riemannian notions in our problem setting, we expect to achieve more efficient and feasible optimization results. The specifics of the RGD implementation and experimental results for our problem will be discussed in the subsequent sections.

B. Analyses of the Landscape and Non-degeneracy

Now we analyze the Karush-Kuhn-Tucker (KKT) conditions for both the original problem and the transformed problem to better understand their relationship.

We first define the following problems:

Original Problem: We first introduce the Lagrangian function, L_O, defined as:

$$L_O(s_i, \mu, \lambda) = ||\Theta(X_{-i}; \theta)s_i - x_i||_2^2 - \mu(1^T s_i - 1) - \lambda^T s_i,$$

where μ and λ are the Lagrange multipliers.

Transformed Problem: For the transformed problem, we define the Lagrangian function as:

$$L_T(z_i, \eta) = ||\Theta(X_{-i}; \theta)(z_i \circ z_i) - x_i||_2^2 - \eta(||z_i||_2^2 - 1),$$

where η is the Lagrange multiplier for the sphere constraint.

With the above definitions, we introduce the following theorems that guarantee the correctness and the non-degeneracy of our solution algorithm.

Theorem 1. *Consider s_i^* as a point fulfilling the second-order Karush-Kuhn-Tucker (KKT) conditions for the original problem. Subsequently, for every z_i^* that adheres to the relation $z_i^* \circ z_i^* = s_i^*$, these points also comply with the second-order KKT conditions within the transformed problem.*

Conversely, if we posit that z_i^* is a point satisfying the second-order KKT conditions for the transformed problem, it follows that $s_i^* = z_i^* \circ z_i^*$ will conform to the second-order KKT conditions as they apply to the original problem.

Theorem 2. *Assume that s_i^* is a non-degenerate second-order KKT point for the original problem. In this case, every z_i^* that follows $z_i^* \circ z_i^* = s_i^*$ acts as a non-degenerate second-order KKT point for the transformed problem.*

Due to space limit, the proofs of the above two theorems are not provided here and they will be provided in the extended journal version of this paper.

C. Perturbed Riemannian Gradient Descent

In this section, we analyze the transformed problem from a Riemannian perspective. Specifically, $g(z_i)$ is seen as a function that operates on the S^{n-1} manifold, thus reformulating the transformed problem into an unconstrained problem in Riemannian optimization.

Definition 1. *Consider z_i^* as a second-order stationary point for the function $g : S^{n-1} \to \mathbb{R}$ when $\nabla g(z_i^*) = 0$ and the smallest eigenvalue of the Hessian matrix, denoted as $\mu_{min}(\nabla^2 g(z_i))$, is greater than or equal to zero. Additionally, we characterize z_i^* as a non-degenerate second-order stationary point of $g : S^{n-1} \to \mathbb{R}$ if $\mu_{min}(\nabla^2 g(z))$ exceeds zero.*

It is interesting to observe that z_i^* is identified as a second-order stationary point in the Riemannian context only if z_i^* also qualifies as a second-order KKT point for the transformed problem [2]. This allows us to use these terms reciprocally, though for the sake of precision, we will persist in using "stationary point" when discussing the transformed problem as a Riemannian optimization problem, and "KKT point" when referring to it as a constrained optimization problem.

Definition 2. *[3] A position z_i on S^{n-1} is characterized as an ϵ-second-order stationary point for the twice-differentiable function $g : S^{n-1} \to \mathbb{R}$ when the following conditions are met: the magnitude of the gradient of g at z_i, $|\nabla g(z_i)|$, does not exceed ϵ and the least eigenvalue of the second derivative of g at z_i, $\mu_{min}(\nabla^2 g(z_i))$, is not less than $-\sqrt{\xi\epsilon}$. Here, ξ is the Lipschitz constant for the Hessian of the 'pullback' of g from the manifold to the tangent space.*

The convergence of Riemannian Gradient Descent (RGD) to a second-order stationary point is not always assured when applied to a nonconvex function - it could potentially arrive at a saddle point. However, prior research [3] has offered optimism by indicating that a perturbed variant of RGD (PGRD) will, in high probability, locate an ϵ-second-order KKT point. We have made modifications to apply this to spherical space in this work.

Theorem 3. *[3] Assume the sequence of iterations, $\{z_k\}_{k=1}^K$, obtained by implementing PRGD on the function $g : S^{n-1} \to \mathbb{R}$ for K iterations. When $K = O\left(\frac{(\log n)^4}{\epsilon^2}\right)$, it is expected that the series $\{z_k\}_{k=1}^K$ will incorporate an ϵ-second-order stationary point of $g(z)$ with a high probability.*

We refer to the combination of the Hadamard parameterization and PRGD as "T-PRGD". Resulting from our analysis of the landscape, we have:

Theorem 4. *Consider the infinite sequence of iterations $\{s_{ik}\}_{k=1}^\infty$ generated by T-PRGD. It is expected that this sequence, $\{s_{ik}\}_{k=1}^\infty$, will contain a subsequence that converges towards a second-order KKT point of the original problem, represented as $s_{ik_\ell} \to s_i^*$, with a high probability.*

We begin by verifying a lemma prior to proceeding with the proof for Theorem 4.

Lemma 3. *Suppose g is twice continuously differentiable. Given any $\delta > 0$, there's a corresponding $\epsilon_\delta > 0$, such that if z_i is considered an ϵ_δ-second-order stationary point, then the inequality $|z_i - z_i^*| < \delta$ holds, where z_i^* denotes a second-order stationary point.*

Proof. Assume $\delta > 0$ and suppose that there is no such ϵ_δ. In such a case, we can select any $\epsilon_k := 1/k$ and identify an ϵ_k-second-order stationary point z_k with the property that $|z_{ik} - z_i^*| \geq \delta$ for all possible second-order stationary points z_i^*.

Given the compactness of S^{n-1}, it's reasonable, if necessary, to presuppose a convergence of z_{ik}: $\lim_{k\to\infty} z_{ik} = \tilde{z}_i \in S^{n-1}$. Through continuity, it follows that $\tilde{z}_i$ is a second-order stationary point:

$$|\nabla g(\tilde{z}_i)| = |\nabla g(\lim_{k\to\infty} z_{ik})| = \lim_{k\to\infty} |\nabla g(z_{ik})| \leq \lim_{k\to\infty} \epsilon_k = 0$$

$$\mu_{\min}(\nabla^2 g(\tilde{z}_i)) = \mu_{\min}(\nabla^2 g(\lim_{k\to\infty} z_{ik})) = \lim_{k\to\infty} \mu_{\min}(\nabla^2 g(z_{ik})) \geq \lim_{k\to\infty} -\sqrt{\xi\epsilon_k} = 0$$

As $\tilde{z}_i = \lim_{k\to\infty} z_{ik}$ this contradicts.

Let's now denote $\{z_{ik}\}_{k=1}^\infty$ as the secondary sequence produced by T-PRGD. Define $\epsilon_\ell := 1/\ell$, and according to Theorem 3, we know that a T_ℓ exists such that $\{z_{ik}\}_{k=1}^{T_\ell}$ contains an ϵ_ℓ-second-order stationary point with high probability (w.h.p.). We will call this point z_{ik_ℓ}. With $\epsilon_\ell \to 0$ and as per the previous lemma, we have $z_{ik_\ell} \to {z_i}^*$, which is a second-order stationary point of the transformed problem. Thus, ${z_i}^*$ coincides with a second-order KKT point of the transformed problem [7]. The sequence $\{s_{ik_\ell} := z_{ik_\ell} \circ z_{ik_\ell}\}_{\ell=1}^\infty$ converges to ${s_i}^* = {z_i}^* \circ {z_i}^*$, which, as per Theorem 1, is a second-order KKT point of the original problem. □

Theorem 5. *Provided that the original problem exhibits the strict saddle property where all local minimizers are also global minimizers, and all second-order KKT points are non-degenerate, then, with high probability, T-PRGD identifies s_{ik} such that the difference between $f(s_{ik})$ and the minimum of*

$f(s_i)$ over Δ^n is less or equal to ϵ, within $K = O((\log n)^4/\epsilon)$ iterations, given that ϵ is sufficiently small.

Proof. We start from a fixed z_{i0}. From Theorem 4, it's established that there is a subsequence $z_{ik\ell} \to {z_i}^*$, which is a second-order stationary point of g, and $s_{ik\ell} \to {s_i}^* := {z_i}^* \circ {z_i}^*$, a second-order KKT point of the original problem. Since the original problem has a strict-saddle property, ${s_i}^*$ is a local minimizer and, as per our assumption, a non-degenerate global minimizer. As per Theorem 2, ${z_i}^*$ is a non-degenerate second-order KKT point of the adjusted problem. As such, ${z_i}^*$ is a non-degenerate second-order stationary point of $g : S^n \to \mathbb{R}$, suggesting that $\nabla^2 g({z_i}^*)$ is positive definite [2]. Given its continuity, a geodesic ball $B({z_i}^*, \delta)$ exists such that $\nabla^2 g(z_i)$ is positive definite for all z_i in $B({z_i}^*, \delta)$. This implies that g confined to $B({z_i}^*, \delta)$ is geodesically τ-strongly convex for some $\tau > 0$, and it complies with the Polyak-Lojasiewicz (PL) condition [2]: $g(z_i) - g({z_i}^*) \leq \frac{1}{2\tau}|\nabla g(z_i)|^2$.

From Lemma 3, we know that an $\epsilon_\delta > 0$ exists such that if z_i is an ϵ_δ-second-order KKT point then $z_i \in B({z_i}^*, \delta)$. Now, let's suppose $\epsilon > 0$ is small enough that $\sqrt{2\tau\epsilon} < \epsilon_\delta$. By Theorem 3, T-PRGD discovers a $\sqrt{2\tau\epsilon}$-second-order KKT point, referred to as $z_{i_{\sqrt{2\tau\epsilon}}}$, within $K = O\left(\frac{(\log n)^4}{(\sqrt{2\tau\epsilon})^2}\right) = O\left(\frac{(\log n)^4}{\epsilon}\right)$ iterations (w.h.p). Since $\sqrt{2\tau\epsilon} < \epsilon_\delta$, we know that $z_{i_{\sqrt{2\tau\epsilon}}}$ is also an ϵ_δ-second-order KKT point, thus $z_{\sqrt{2\tau\epsilon}} \in B({z_i}^*, \delta)$. Utilizing the PL condition and defining $s_{i_{\sqrt{2\tau\epsilon}}} = z_{i_{\sqrt{2\tau\epsilon}}} \circ z_{i_{\sqrt{2\tau\epsilon}}}$, we can write:

$$f(s_{i_{\sqrt{2\tau\epsilon}}}) - \min_{s_i \in \Delta^n} f(s_i) = f(s_{i_{\sqrt{2\tau\epsilon}}}) - f(s_i^*)$$

$$= g(z_{i_{\sqrt{2\tau\epsilon}}}) - g(z_i^*) \leq \frac{1}{2\tau}\|\nabla g(z_{i_{\sqrt{2\tau\epsilon}}})\|^2 \leq \frac{1}{2\tau}(\sqrt{2\tau\epsilon})^2 \leq \epsilon$$

□

IV. Experimental Evaluation

To validate the effectiveness and efficiency of our proposed method, we conducted comprehensive experiments. In this section, we introduce the benchmark datasets used, describe the baseline methods for comparison, outline the experimental setup, present the evaluation metrics used, and discuss the results obtained.

A. Datasets

We selected a diverse set of benchmark datasets to evaluate the robustness and versatility of our method. Table I provides a summary of the datasets used, including the number of instances, features, and classes.

The selected datasets cover various domains, including facial images (Olivetti), object images (COIL-20), and citation networks (Cora). These datasets served as the foundation for our experiments. To assess the algorithm's robustness, we introduced perturbations at a 30% noise level. For image datasets (Olivetti and COIL-20), Gaussian noise was added to pixel values using a perturbation-dependent standard deviation. For the Cora datasets, representing citation networks and biomedical literature, respectively, we randomly shuffled a small percentage of node/document attributes to induce perturbations. This controlled noise injection ensured that the perturbations closely resembled real-world variations.

TABLE I: Description of benchmark datasets used in the experiments.

Dataset	Instances	Features	Classes
Olivetti	400	4096	40
COIL-20	1440	1024	20
Cora	2708	1433	7

B. Comparative Methods and Evaluation Metrics

We compare our proposed method against state-of-the-art approaches in three categories:

Category 1: Adjacency Matrix-based Methods: Representatives include Spectral Clustering (SC) and NetMF [9]. These methods generate node embeddings solely based on the adjacency matrix and employ k-means clustering.

Category 2: Direct Embedding-based Methods: Representatives include AutoEncoder (AE) and NMF. These methods directly generate lower-dimensional embeddings and utilize k-means clustering as a post-processing step.

Category 3: End-to-End Graph Neural Network (GNN) Models: Representatives include DiffPool [17] and MinCut [1]. These GNN models generate soft cluster assignments by considering both graph connectivity and node features.

In our work, we leverage Graph Isomorphism Networks (GIN) [16] as the deep representation. GIN employs multiple graph convolutional layers with non-linear activation functions to capture information from larger neighborhood scopes. This enables GIN to effectively capture graph structure and learn expressive node embeddings. One key advantage of our proposed method, BSSR, is its parameter-independent nature, eliminating the need for extensive parameter tuning. This allows us to achieve reliable and accurate results without the burden of parameter optimization.

We evaluate our method using a 5-fold cross-validation approach, repeated ten times for robust statistical analysis. We utilize Average Accuracy (ACC) and Normalized Mutual Information (NMI) as evaluation metrics, assessing the correctness of predicted cluster assignments and the agreement between predicted clusters and ground truth labels.

Furthermore, we compare our proposed optimization method, Tangent Perturbed Riemannian Gradient Descent (T-PRGD), against the Projected Gradient Descent method (PGD) to demonstrate its superior performance in achieving accurate and reliable graph clustering results.

C. Empirical Studies of the Convergency of the Proposed Algorithm

The findings from Figure 1 offer valuable insights into the performance of our proposed optimization method, T-PRGD, compared to the traditional method, PGD, on the Olivetti dataset.

Firstly, the figure demonstrates that T-PRGD achieves faster convergence than PGD. The steeper decreasing rate of the loss curve for T-PRGD indicates that our method reaches lower loss

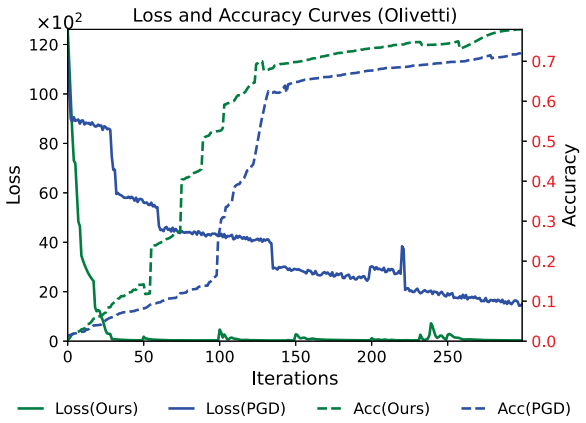

Fig. 1: Loss and Accuracy Comparison of T-PRGD and PGD on the Olivetti dataset. (Green line: T-PRGD, Blue line: PGD. Solid line: Loss, Dashed line: Accuracy.)

values in fewer iterations. This faster convergence is essential for efficient optimization and can result in significant time savings in real-world applications. Furthermore, the figure highlights the superior capability of T-PRGD in overcoming saddle points. The presence of "plateaus" in the loss curves suggests the existence of saddle points, where the gradient becomes close to zero and conventional optimization algorithms tend to stagnate. In contrast, T-PRGD exhibits a fluctuating loss curve, indicating its ability to explore alternative directions and avoid getting trapped in suboptimal solutions.

Importantly, these findings align with our previous theoretical analysis, which emphasized T-PRGD's potential to converge to a global optimum and effectively navigate saddle points. The observed lower loss achieved by T-PRGD further supports these theoretical claims. Moreover, the accuracy (ACC) curve of T-PRGD outperforms that of PGD, indicating superior clustering performance. This demonstrates the effectiveness of T-PRGD in producing more accurate data similarity measures, aligning with our expectations.

Overall, the empirical evidence from Figure 1 confirms the superiority of our proposed method, T-PRGD. It exhibits faster convergence, effectively overcomes saddle points, and achieves improved accuracy. These findings align with our theoretical analysis, showcasing the practical advantages of employing T-PRGD for data similarity measurement.

D. Comparative Studies of the Proposed Algorithm in Clustering Tasks

In Table II and Table III, the comparative analysis of different methods on the original and perturbed datasets provides compelling evidence for the effectiveness and superiority of our proposed method, BSSR-GIN, when compared to state-of-the-art approaches.

Across the original datasets, BSSR-GIN consistently achieves superior performance in terms of average accuracy score (Acc score) and normalized mutual information (NMI), confirming its ability to accurately identify data similarity and capture the underlying structure of the data. This exceptional performance can be attributed to the robustness and scalability of BSSR-GIN, leveraging the power of GIN to capture complex graph structures and learn expressive node embeddings, as well as the advanced parameter-independent BSSR approach. These characteristics enable BSSR-GIN to effectively handle scale inconsistencies and outlier noise, leading to more reliable and precise cluster assignments. Furthermore, the integration of shift-invariant sparse representation in BSSR-GIN further enhances its clustering performance. By incorporating a simplex constraint, our method promotes sparser representations while ensuring shift-invariance, leading to more accurate and interpretable clustering results. This unique design choice enables a precise analysis of the data, providing deeper insights into its underlying patterns.

Overall, the experimental results showcased in the tables strongly validate the advantages and motivations of our novel method, BSSR-GIN. Its robustness, scalability, shift-invariant sparse representation, and optimized optimization process all contribute to its exceptional performance. By outperforming other state-of-the-art methods in terms of average accuracy score and normalized mutual information, BSSR-GIN demonstrates its effectiveness in accurately uncovering the intrinsic structure of complex datasets and providing valuable insights for a wide range of data analysis tasks.

V. Conclusion

We propose the Beyond Simplex Sparse Representation (BSSR) method, which effectively constructs reliable and interpretable graphical representations of high-dimensional data. By leveraging sparse representation and deep learning techniques, BSSR addresses challenges such as parameter dependencies and scale inconsistencies. We also introduce the Tangent Perturbed Riemannian Gradient Descent (T-PRGD) algorithm, which optimizes the standard simplex by transforming the constraint onto the Riemannian manifold of the unit sphere. Experimental evaluation demonstrates the superiority of our methods compared to state-of-the-art approaches, showcasing their accuracy and robustness. Our work contributes significant advancements to data representation and optimization, enabling more accurate analysis in the era of big data.

Acknowledgment

Corresponding author: Hua Wang (huawangcs@gmail.com).

This work was supported in part by the National Science Foundation (NSF) under the grants of IIS 1652943, IIS 1849359, CNS 1932482.

References

[1] Filippo Maria Bianchi, Daniele Grattarola, and Cesare Alippi. Spectral clustering with graph neural networks for graph pooling. In *International conference on machine learning*, pages 874–883. PMLR, 2020.

[2] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

[3] Christopher Criscitiello and Nicolas Boumal. Efficiently escaping saddle points on manifolds. *Advances in Neural Information Processing Systems*, 32, 2019.

TABLE II: Clustering performance of different methods on the original datasets and the perturbed datasets with 30% noise, , including the Olivetti, COIL-20, and Cora datasets. The evaluation metric is the average accuracy score (Acc score) with their standard deviations.

Dataset	SC	NetMF	AE	NMF
Olivetti	0.598 ± 0.054	0.498 ± 0.079	0.601 ± 0.056	0.662 ± 0.052
COIL-20	0.401 ± 0.002	0.404 ± 0.010	0.477 ± 0.061	0.521 ± 0.044
Cora	0.484 ± 0.035	0.488 ± 0.025	0.515 ± 0.034	0.432 ± 0.039
Olivetti (30% noise)	0.345 ± 0.042	0.361 ± 0.053	0.373 ± 0.047	0.210 ± 0.039
COIL-20 (30% noise)	0.201 ± 0.007	0.281 ± 0.014	0.292 ± 0.044	0.186 ± 0.035
Cora (30% noise)	0.269 ± 0.029	0.290 ± 0.020	0.288 ± 0.027	0.191 ± 0.032
Dataset	**DiffPool**	**MinCut**	**GIN**	**BSSR-GIN**
Olivetti	0.608 ± 0.032	0.604 ± 0.044	0.610 ± 0.044	**0.788 ± 0.023**
COIL-20	0.520 ± 0.027	0.516 ± 0.004	0.507 ± 0.004	**0.638 ± 0.033**
Cora	0.621 ± 0.036	0.523 ± 0.036	0.620 ± 0.031	**0.625 ± 0.030**
Olivetti (30% noise)	0.543 ± 0.032	0.539 ± 0.039	0.548 ± 0.038	**0.585 ± 0.025**
COIL-20 (30% noise)	0.403 ± 0.024	0.397 ± 0.005	0.421 ± 0.007	**0.462 ± 0.027**
Cora (30% noise)	0.494 ± 0.031	0.502 ± 0.030	0.499 ± 0.027	**0.507 ± 0.025**

TABLE III: Clustering performance of different methods on the original datasets and the perturbed datasets with 30% noise, including the Olivetti, COIL-20, and Cora datasets. The evaluation metric is the normalized mutual information (NMI) with their standard deviations.

Dataset	SC	NetMF	AE	NMF
Olivetti	0.652 ± 0.045	0.564 ± 0.051	0.612 ± 0.038	0.571 ± 0.024
COIL-20	0.518 ± 0.011	0.522 ± 0.038	0.579 ± 0.048	0.531 ± 0.036
Cora	0.592 ± 0.030	0.597 ± 0.027	0.625 ± 0.033	0.547 ± 0.037
Olivetti (30% noise)	0.352 ± 0.049	0.372 ± 0.068	0.381 ± 0.051	0.228 ± 0.043
COIL-20 (30% noise)	0.211 ± 0.008	0.294 ± 0.015	0.307 ± 0.048	0.201 ± 0.037
Cora (30% noise)	0.284 ± 0.031	0.304 ± 0.021	0.304 ± 0.029	0.202 ± 0.034
Dataset	**DiffPool**	**MinCut**	**GIN**	**BSSR-GIN**
Olivetti	0.598 ± 0.028	0.594 ± 0.042	0.602 ± 0.038	**0.792 ± 0.021**
COIL-20	0.530 ± 0.032	0.526 ± 0.005	0.515 ± 0.004	**0.646 ± 0.029**
Cora	0.701 ± 0.032	0.600 ± 0.034	0.680 ± 0.029	**0.705 ± 0.026**
Olivetti (30% noise)	0.558 ± 0.035	0.554 ± 0.041	0.560 ± 0.042	**0.598 ± 0.028**
COIL-20 (30% noise)	0.420 ± 0.025	0.416 ± 0.006	0.437 ± 0.007	**0.483 ± 0.029**
Cora (30% noise)	0.511 ± 0.033	0.518 ± 0.032	0.515 ± 0.029	**0.525 ± 0.027**

[4] Wei Jin, Tyler Derr, Yiqi Wang, Yao Ma, Zitao Liu, and Jiliang Tang. Node similarity preserving graph convolutional networks. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 148–156, 2021.

[5] Zhao Kang, Guoxin Shi, Shudong Huang, Wenyu Chen, Xiaorong Pu, Joey Tianyi Zhou, and Zenglin Xu. Multi-graph fusion for multi-view spectral clustering. *Knowledge-Based Systems*, 189:105102, 2020.

[6] Mayank Kejriwal. *Domain-specific knowledge graph construction*. Springer, 2019.

[7] David G Luenberger. The gradient projection method along geodesics. *Management Science*, 18(11):620–631, 1972.

[8] Feiping Nie, Hua Wang, Heng Huang, and Chris Ding. Early active learning via robust representation and structured sparsity. In *Twenty-third international joint conference on artificial intelligence*, 2013.

[9] Jiezhong Qiu, Yuxiao Dong, Hao Ma, Jian Li, Kuansan Wang, and Jie Tang. Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 459–467, 2018.

[10] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[11] Tomas Vaskevicius, Varun Kanade, and Patrick Rebeschini. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32, 2019.

[12] Hua Wang, Feiping Nie, and Heng Huang. Multi-view clustering and feature learning via structured sparsity. In *International conference on machine learning*, pages 352–360. PMLR, 2013.

[13] Hua Wang, Feiping Nie, and Heng Huang. Robust and discriminative self-taught learning. In *International conference on machine learning*, pages 298–306. PMLR, 2013.

[14] Hua Wang, Feiping Nie, Heng Huang, and Chris Ding. Heterogeneous visual features fusion via sparse multimodal machine. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3097–3102, 2013.

[15] Hua Wang, Feiping Nie, Heng Huang, Shannon Risacher, Chris Ding, Andrew J Saykin, and Li Shen. Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance. In *2011 International Conference on Computer Vision*, pages 557–562. IEEE, 2011.

[16] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[17] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. *Advances in neural information processing systems*, 31, 2018.

[18] Chong You, Zhihui Zhu, Qing Qu, and Yi Ma. Robust recovery via implicit bias of discrepant learning rates for double over-parameterization. *NIPS*, 33:17733–17744, 2020.