Sparsity on Statistical Simplexes and Diversity in Social Ranking

SUN, Ke, MOHAMED, Hisham, MARCHAND-Maillet, Stéphane

Abstract

Sparsity in R^m has been widely explored in machine learning. We study sparsity on a statistical simplex consisting of all categorical distributions. This is different from the case in R^m because such a simplex is a Riemannian manifold, a curved space. A learner with sparse constraints should be likely to fall to its low-dimensional boundaries. We present a novel analysis on the statistical simplex as a manifold with boundary. The main contribution is an explicit view of the learning dynamics in between high-dimensional models in the interior of the simplex and low-dimensional models on its boundaries. We prove the differentiability of the cost function, the natural gradient with respect to the Riemannian structure, and convexity around the singular regions. We uncover an interesting relationship with L1 regularization. We apply the proposed technique to social network analysis. Given a directed graph, the task is to rank a subset of influencer nodes. Here, sparsity means that the top-ranked nodes should present diversity in the sense of minimizing influence overlap. We present a ranking algorithm based on the natural [...]
Sparsity on Statistical Simplexes and Diversity in Social Ranking

Ke Sun
Hisham Mohamed
Stéphane Marchand-Maillet

Viper Group, Computer Vision and Multimedia Laboratory, University of Geneva

Abstract

We study sparsity on a statistical simplex consisting of all categorical distributions. This is different from the case in $\mathbb{R}^m$ because such a simplex is a Riemannian manifold, a curved space. A learner with sparse constraints would be likely to fall into its low-dimensional boundaries. We present a novel analysis on the statistical simplex as a manifold with boundary. We investigate the learning dynamics in between high-dimensional models in the interior of the simplex and low-dimensional models on its boundaries. We study the differentiability of the cost function and its natural gradient with respect to the Riemannian structure.

We apply the proposed technique to social network analysis. Given a directed graph, the task is to rank a subset of influencer nodes. Here, sparsity means that the top-ranked nodes should present diversity in the sense of minimizing influence overlap. We present a ranking algorithm based on the natural gradient. It can scale up to graph datasets with millions of nodes. On real large networks, its top-ranked nodes are the most influential among several commonly-used techniques.

Keywords: Sparsity, Ranking, Information Geometry

1. Introduction

Sparsity has been a main topic of machine learning (Tibshirani, 1996; Ng, 2004; Zhao and Yu, 2006; Bach, 2008). The majority of previous works concentrated on studying sparsity in an Euclidean space, where a model parameter $\alpha \in \mathbb{R}^n$ is constrained to be likely on certain subspaces of $\mathbb{R}^n$ through $L_1$-type regularization or, equivalently, a Laplace prior distribution of $\alpha$. This results in a “simple” model, in the sense that only a few entries of $\alpha$ are non-zero.

Recently, the notion of sparsity has been extended (Pilanci et al., 2012; Kyrillidis et al., 2012) to the statistical simplex $\mathbb{S}^m = \{(\eta_1, \ldots, \eta_m) : \forall j, \eta_j \geq 0; \sum_{j=1}^m \eta_j \leq 1\}$, meaning that only a small number of $\eta_j$’s are non-zero. $L_1$-norm-based techniques are not ideal because (1) $L_1$ norm depends on the coordinate system and thus is not an intrinsic measure; (2) $L_1$ norm in the $\eta$-coordinates already appears as a constraint. Pilanci et al. (2012) proposed a relaxation of the minimization problem on the number of non-zero $\eta_j$’s. Kyrillidis et al. (2012) studied sparsity based on an Euclidean projection onto some sparse region.

1. The upper script of a manifold, e.g. “m” in “$\mathbb{S}^m$”, denotes the dimensionality.
Sparsity on Statistical Simplexes

on $S^m$. In these methods, $S^m$ is studied as a subset of the ambient $R^{m+1}$, and sparsity is derived from the Euclidean geometry. However, in many cases, $\eta \in S^m$ means a probability distribution. $S^m$ is not Euclidean but instead has a unique information geometry (Rao, 1945; Čencov, 1982; Amari and Nagaoka, 2000). To understand sparsity in such a geometric way and to study sparsity that is invariant under re-parametrization is of theoretical interest.

Based on existing methods, we present an information geometric analysis on the statistical simplex as a manifold with boundary. This is novel because past efforts mainly focused on the interior of $S^m$. While a learner can jump in-between the boundary $\partial S^m$ and inside $S^m$ (Ghahramani and Beal, 2000; Xu, 2009), the learning dynamics near $\partial S^m$ are not explicitly investigated. We discovered that the learning cost function is decomposed into a smooth term and a non-smooth term, where smoothness is defined on the manifold with boundary. The non-smooth term, as a coordinate-invariant regularization, helps to create singularities near $\partial S^m$, where the gradient flow is always inward, i.e., from $\partial S^m$ to the interior of $S^m$.

As an applicative contribution, we investigate such sparsity in graph-based ranking (Page et al., 1999). The task is to rank a subset of nodes in a social network, so that they can maximally spread influence. This is reduced to inferring a probability distribution on the graph nodes. Sparsity in this context means that a limited number of nodes have a non-zero probability of being an influencer. This agrees with recent interests in graph-based information retrieval to retrieve a diversity of nodes (Zhu et al., 2007). We propose a scalable implementation along with a novel usage of natural gradient (Amari, 1998). Through experimenting on real large networks, we show that the proposed ranking most effectively discovers important nodes to maximally cover the network.

The rest of this paper is organized as follows. Section 2 introduces some prerequisites of information geometry, then presents an analysis on sparsity on statistical simplexes. The theoretical results (theorems 3 and 4) are in subsection 2.2. Section 3 discusses an application on social network ranking and the associated learning algorithm. Section 4 discusses related works and compares the proposed ranking with PageRank. Section 5 presents an experimental study on real large networks. Section 6 concludes and discusses possible extensions.

2. Sparsity on Statistical Simplexes

An observable random variable $X$, either discrete or continuous, is associated with a latent random binary vector $Y \in \left\{ (y_1, \ldots, y_m) : \forall j, y_j = 0 \text{ or } 1; \sum_{j=1}^{m} y_j \leq 1 \right\}$, with at most one bit equal to “1” and following a discrete distribution. We assume that $p(X \mid Y)$ is given. This simplified case lets us focus on the central issue, i.e. sparsity, and is useful for social network analysis (to be introduced in section 3). Based on a set of independent and identically distributed observations (i.i.d.) $X = \{ X^1, \ldots, X^n \}$, the problem is to infer a prior distribution $p(Y)$, where $Y$ should be sparse. Sparsity means that only a small subset of bits in $Y$ are activated, i.e., having a non-zero probability of being “1”, while the rest bits are deactivated, i.e., always “0”.

17
2.1. Prerequisites

$p(Y)$ is in the exponential family of distributions, as it can be written in the canonical form

$$p(Y | \theta) = \exp \left( \sum_{j=1}^{m} \theta_j y_j - \psi(\theta) \right),$$

where $\theta = (\theta_1, \ldots, \theta_m)$ are the canonical parameters ranging in $\mathbb{R}^m$, and $\psi(\theta) = \log (1 + \sum_{j=1}^{m} \exp \theta_j)$. Another way of representing $p(Y)$ is by $p(Y | \eta) = \sum_{j=1}^{m} y_j \eta_j + (1 - \sum_{j=1}^{m} y_j) \eta_0$, where $\eta = (\eta_1, \ldots, \eta_m)$ are the expectation parameters ranging in $S^m - \partial S^m$, and $\eta_0 = 1 - \sum_{j=1}^{m} \eta_j$. The $\theta$- and $\eta$-coordinate systems relate with each other by the Legendre transformations (Amari and Nagaoka, 2000)

$$\frac{\partial \psi}{\partial \theta} = \eta, \quad \frac{\partial \varphi}{\partial \eta} = \theta,$$

where $\varphi(\eta) = \sum_{j=0}^{m} \eta_j \log \eta_j$ is the negative entropy. Both $\psi(\theta)$ and $\varphi(\eta)$ are strictly convex functions with respect to $\theta \in \mathbb{R}^m$ and $\eta \in S^m$, respectively.

The statistical manifold consisting of all such $p(Y)$ is equipped with a Riemannian metric (Lee, 2012), which can be intuitively understood as a local inner product defined on each point and varying smoothly along the manifold. It was showed (Rao, 1945; Čencov, 1982) that Fisher Information Metric (FIM) $g_{ij}(\theta) = -E(\partial^2 \log p/\partial \theta^2)$ is the unique Riemannian metric under some conditions, where $E(\cdot)$ is the expectation with respect to $p(Y | \theta)$. By eqs. (1) and (2), $g_{ij}(\theta) = \partial^2 \psi/\partial \theta^2 = \partial \eta/\partial \theta$ coincides with the Jacobi matrix $\partial \eta/\partial \theta$ of the coordinate transformation $\theta \rightarrow \eta$. Similarly, FIM with respect to the $\eta$-coordinates is $g_{ij}(\eta) = \partial^2 \varphi/\partial \eta^2 = \partial \theta/\partial \eta$, which is the inverse of $g_{ij}(\theta)$. It can be verified that $g(\theta)$ and $g(\eta)$ are essentially the same metric by showing $\langle a \partial \theta, b \partial \theta \rangle_{g(\theta)} = \langle \partial \eta/\partial \theta a \partial \eta, \partial \eta/\partial \theta b \partial \eta \rangle_{g(\eta)}$. $a \partial \theta$ denotes the vector field (Lee, 2012) $\sum_{j=1}^{m} a_j \partial \theta_j$, which can be understood as real vectors in local linearizations of the Riemannian manifold $S^m$. $\langle \cdot, \cdot \rangle_{g(\theta)}$ denotes the inner product with respect to the Riemannian metric $g(\theta)$.

In this paper, FIM is used to compute the natural gradient (Amari and Nagaoka, 2000), i.e. the gradient with respect to the Riemannian geometry, of a smooth function $f$ on $S^m$. By definition, the natural gradient of $f$ is $\text{grad} f = (g_{ij}(\theta))^{-1} \partial f/\partial \theta \cdot \partial \theta = \partial f/\partial \eta \cdot \partial \eta$.

$\text{grad} f$ is invariant to the choice of the coordinate system. For example, with respect to the $\eta$-coordinates, $\text{grad} f = (g_{ij}(\eta))^{-1} \partial f/\partial \eta \cdot \partial \eta$ is exactly the same $\text{grad} f$ up to coordinate transformation.

Maximum-likelihood learning can be implemented (Amari, 1995) by

$$\begin{cases} \theta^i = \theta + c^i, \quad \forall i = 1, \ldots, n; \\ \tilde{\eta} = \sum_{i=1}^{n} \eta^i/n; \\ \min_{\theta} E_r(\theta), \quad E_r(\theta) = \psi(\theta) + \tau \varphi(\tilde{\eta}) - \theta^T \tilde{\eta}. \end{cases}$$

The first equation is the Bayes’ rule, where $c^i_j = \log p(X^i | Y_j = 1) - \log p(X^i | Y = 0)$, and $\theta^i$ denotes the posterior estimation with regard to $X^i$. The second equation summarizes the posterior estimations into a new $\tilde{\eta}$ (or $\tilde{\theta}$). The last line in eq. (3) minimizes the difference

2. One can re-write $\theta^i = \theta + c^i$ in the canonical form of Bayes’ rule in the $\eta$-coordinates.
Sparsity on Statistical Simplexes

\( f = C_{\text{small}} \)

\( f = C_{\text{large}} \)

Figure 1: (a) The low-entropy region \( \{ \eta : -\varphi(\eta) \leq C \} \) in \( S^2 \), where \( C \approx 0.6 \); (b) A constrained optimization \( \min f(\eta) \), s.t. \( -\varphi(\eta) \leq C \). \( f(\eta) \) is showed by the blue contours.

between \( \theta \) and its image \( \tilde{\eta} \) after the first and second equations, under the intuition that a locally optimal \( \theta \) should coincide with \( \tilde{\eta} \). Note, \( E_1 \) is exactly the Kullback-Leibler (KL) divergence from \( \tilde{\eta} \) to \( \theta \). This learning is supported by the following propositions.

**Proposition 1**

\[ E_1(\theta) \geq 0; \quad E_1(\theta) = 0 \text{ if and only if } \theta \text{ is a stationary point of the log-likelihood function } L(\theta) = \sum_{i=1}^{n} \log p(X^i | \theta). \]

Proposition 1 says that \( E_1(\theta) \) is an “indicator function”, reaching its minimum at, and only at, the stationary points of \( L(\theta) \). The proof is straightforward by writing \( L(\theta) = \sum_{i=1}^{n} \log \sum_{Y} (p(Y|\theta) p(X^i|Y)) \) and computing its differential according to eq. (1).

**Proposition 2**

(1) \( \text{grad} \ L = n(\tilde{\eta} - \eta) \partial \eta; \) (2) \( \text{grad} \ E_\tau = \left( \eta - \tilde{\eta} + \frac{\partial \eta}{\partial \theta} (\tau \tilde{\theta} - \theta) \right) \partial \eta. \)

Proposition 2, which can be derived from the definition of the natural gradient introduced earlier and eqs. (2) and (3), gives the natural gradients of \( L(\theta) \) and \( E_\tau(\theta) \). It shows that minimizing \( E_1(\theta) \) instead of maximizing \( L(\theta) \) benefits from another gradient term pulling together \( \theta \) and \( \tilde{\theta} \). This explains the faster convergence of learning as a two-body problem (Amari, 1995). Here, the two bodies \( \theta \) and \( \tilde{\eta} \), which are marginal distributions \( p(Y) \), and the learning gradient flow are all in one simple space \( S^m \). Equation (3) is just one (representative) method in a widely-studied spectrum. Therefore, the information geometric analysis in the following subsection 2.2 could be useful in more general contexts.

### 2.2. Sparsity on Statistical Simplexes

The \( \eta \)-coordinates expose a hierarchy of statistical manifolds. This allows us study singular regions (Amari et al., 2006) and impose sparsity on \( Y \). The closed simplex \( S^m \) is a manifold with boundary (Lee, 2012), where any point has a neighborhood which is like an open subset of \( R^m_+ = \{ \alpha \in \mathbb{R}^m : \alpha_1 \geq 0 \} \). There are “corners” in \( S^m \), which do not satisfy such a property. They are less interesting and will be ignored in subsequent discussions. A face of \( S^m \) is a statistical manifold with exactly the same structure as \( S^m \) but one less dimension, corresponding to the sparse case where some bit of \( Y \) is deactivated. This paper considers the learning dynamics on \( S^m \) as a whole without excluding \( \partial S^m \). A learner can go from inside \( S^m \) to \( \partial S^m \) or the other way round.
A natural idea to make $Y$ sparse is to penalize the entropy $-\varphi$ by setting $\tau < 1$ in eq. (3). This gives an intrinsic regularization as the entropy is invariant to the choice of the coordinate system. This can be understood by the fact that $\varphi(\eta)$ is the KL-divergence from $\eta$ to the simplex center plus some constant. To gain some intuitions, consider $S^2$. Figure 1(a) shows the low entropy region $\{\eta : -\varphi(\eta) \leq C\}$, where $C \approx 0.6$. Such a region has some sharp corners, which can easily trap the optimizer of a constrained problem $\min f(\eta)$, s.t. $-\varphi(\eta) \leq C$, where $f(\eta)$ is a smooth function on $S^2$. This is only an intuitive view in the $\eta$-coordinate system. Formally, we have to consider the smoothness of the learning cost function $E_\tau$ in eq. (3). A smooth function $f$ on the manifold with boundary $S^m$ means that the differentials of $f$ continuously extend to an open neighbourhood of $\eta$ in $\mathbb{R}^m$ for any $\eta \in S^m$. Even if $\eta$ is a bit outside $S^m$, these differentials are still well-defined. We have the following result 3.

**Theorem 3** Assume that $\forall j, \exists i$, s.t. $p(X^i | Y_j = 1) > 0$. Then, (1) $E_1$ is a smooth function on $\mathcal{F} = \{\eta \in S^m | \forall i, p(X^i | \eta) > 0\}$; (2) $\forall \tau < 1$, $E_\tau$ is continuous on $\mathcal{F}$ but non-differentiable on $\{\eta \in \mathcal{F} : \exists j, \eta_j = 0; \forall i \neq j, \eta_i > 0\}$.

In the above theorem 3, the assumption means that any bit in $Y$ is associated to at least one observation $X^i$, otherwise it can be removed without affecting the system. $\mathcal{F}$ is a feasible region consisting of all such $\eta$ that “covers” all observations. If there is some redundant bit(s) in $Y$ that can be deactivated, then $\mathcal{F} \cap \partial S^m \neq \emptyset$. The smoothness of $E_1$ does not rely on the choice of the coordinate system and therefore reflects an intrinsic property. A learner based on $E_1$ is “unaware” of $\partial S^m$, meaning that it does not treat $\partial S^m$ in a particular way. During learning, it could go from inside $S^m$ to $\partial S^m$ or the other way round. Often, it tends to go from $\partial S^m$ to inside $S^m$, because a model inside $S^m$ has a higher complexity and a higher potential likelihood.

By making $\tau$ smaller than 1, $E_\tau$ becomes non-differentiable on $\partial S^m$. This reveals an interesting relationship with sparsity on $\mathbb{R}^m$ (Ng, 2004). $L_1$ norm, which is non-differentiable on $\{\alpha \in \mathbb{R}^m : \exists j, \alpha_j = 0\}$, is used to enforce sparsity on $\mathbb{R}^m$. The entropy function $-\varphi$, which is non-differentiable on $\partial S^m$, is used to enforce sparsity on $S^m$. The level sets of $\varphi$ have the form $d\varphi = \sum_{j=1}^m \partial \varphi / \partial \eta_j d\eta_j = \sum_{j=1}^m \theta_j d\eta_j = 0$. This differential representation is similar to the level sets of $L_2$ norm in an Euclidean space. This helps to understand why $\varphi$ plays a similar role of a norm. Theorem 3 tells that the surface of $E_\tau$ is singular on $\partial S^m$, but it does not describe its gradient flow near $\partial S^m$ inside $S^m$. This is covered by the following theorem.

**Theorem 4** If $\tau < 1$, $\langle \text{grad} E_\tau, \partial / \partial \eta_j \rangle_g \to \infty$ as $\eta_j \to 0^+$ inside $\mathcal{F}$.

**Remark 5** Although the statement is based on $\eta$-coordinates, the natural gradient $\text{grad} E_\tau$ is invariant to the coordinate system. In another coordinate system $\zeta$, $\text{grad} E_\tau$ is still “inward”, meaning it flows from $\partial S^m$ to inside $S^m$.

As we are minimizing $E_\tau$ in eq. (3), learning is along the vector field $-\text{grad} E_\tau$. Theorem 4 says that whenever the learner approaches $\partial S^m$ from inside $S^m$, and $\eta_j$ becomes small enough, it will go to $\partial S^m$. Therefore, there is a continuous “attractive region” near $\partial S^m$.

where any $\boldsymbol{\eta}$ will be pulled into $\partial S^{m}$. The size of this region depends on $\tau$. If $\tau$ is slightly smaller than 1, the surface of $E_{\tau}$ is only bent down near $\partial S^{m}$. As $\tau$ turns smaller, the attractive region widens. This also means that a learner can only reach $\partial S^{m}$ from inside $S^{m}$ but not go from $\partial S^{m}$ to inside $S^{m}$. From an algorithmic perspective, the problem scale is reduced during optimization. Learning becomes more and more efficient. On the other hand, any local optimum solution on $\partial S^{m}$ also has such attraction to a leaner inside $S^{m}$. If the learner falls into $\partial S^{m}$ and deactivates a bit of $Y$ that is in the global optimal solution, there is no way to reverse it in subsequent learning. The optimization must carefully explore the feasible region inside $S^{m}$ before falling into $\partial S^{m}$.

Consider $E_{\tau}$ as a function of $(\boldsymbol{\theta}, \boldsymbol{\eta}) \in S^{m} \times S^{m}$ and the case $0 \leq \tau < 1$. By $\partial E_{\tau}/\partial \boldsymbol{\eta} = 0$, we get $\tau \hat{\theta} = \boldsymbol{\theta}$. As the absolute value of $\theta$ is large near $\partial S^{m}$, $\min E_{\tau}$ causes an isotropic scaling $\theta \rightarrow \boldsymbol{\eta}$ towards $\partial S^{m}$ in the $\theta$-coordinate system. From a second-order view, the Hessian of $E_{\tau}$ is

$$H = \begin{bmatrix} g(\boldsymbol{\theta}) & -I \\ -I & \tau g(\boldsymbol{\eta}) \end{bmatrix},$$

where $g$ is FIM. $E_{\tau}$ is convex with respect to $\theta$ and $\boldsymbol{\eta}$ individually. The joint convexity is guaranteed if and only if $\tau g(\boldsymbol{\eta}) \succeq g^{-1}(\boldsymbol{\theta})$, which is equivalent to $\tau g(\boldsymbol{\eta}) \geq g(\boldsymbol{\eta})$. By simple derivations, $|g(\boldsymbol{\eta})|$ turns large as $\eta$ moves from the simplex center to $\partial S^{m}$. Therefore, the joint convexity of $E_{\tau}$ means that $\boldsymbol{\eta}$ should be closer to $\partial S^{m}$ as compared to $\theta$. This scaling in the $\theta$-coordinates forms a mechanism, making the two-body system to be likely to reach $\partial S^{m}$. Note, we do not consider the case $\tau < 0$, when $E_{\tau}$ is concave with respect to $\boldsymbol{\eta}$ and easily causes trivial solutions without strong constraints.

### 3. Application to Graph-based Ranking

Consider a social network given by a directed graph $G = (V; E)$, where $V$ are the nodes, and $E \subset \{(i, j) : i \in V; j \in V\}$ are the links. Each link $(i, j) \in E$ from node $i$ to node $j$ is associated with a weight $w_{ij} > 0$. Usually, $(i, j)$ means that $j$ can influence $i$ with the strength $w_{ij}$. For example, in twitter, $(i, j)$ means that $i$ reads micro-blogs posted by the individual $j$; in citation networks, $(i, j)$ means that the article $i$ is based on a previous article $j$. A subset $V_{I} \subset V$ of size $|V_{I}| = m$, referred to as the influences, are considered as potential candidates to emit influence. Their states of being chosen can be represented as an $m$-dimensional latent binary vector $Y$ as in section 2. Without loss of generality, we set $V_{I}$ to be the set of nodes with at least one incoming link. On the other hand, a target population $V_{O} \subset V$, characterized by the random variable $X$, plays the role of receiving influence. By default, we set $V_{O}$ to be all nodes in $V$ with at least one out-going link. An influencer in $V_{I}$ indexed by $j$ ($1 \leq j \leq m$) can influence any node in $V_{O}$ indexed by $i$ ($1 \leq i \leq n$) according to a given influence matrix $F_{m \times n}$. $F$ gives the conditional distribution $p(X \mid Y)$ such that $\forall j, \forall i, F_{ji} \geq 0$ and $\forall j, \sum_{i=1}^{n} F_{ji} = 1$. $F$ is usually sparse, pre-computed according to the graph structure and the link weights. We let

$$F_{ji} = \begin{cases} \frac{w_{ij}}{\sum_{i:i \rightarrow j} w_{ij}} & \text{if } (i, j) \in E; \\ 0 & \text{if otherwise,} \end{cases}$$

meaning that an influencer $j$ influences each of its predecessors with a strength proportional to the link weight. Based on the modeling in section 2, a prior distribution $\boldsymbol{\eta} = (\eta_{1}, \ldots, \eta_{n})$
can be learned, where for any influencer , this can be used to rank the influencers while presenting diversity in the ranking results.

**Ranking** Consider an information diffusion process in a social network . A piece of information, e.g. marketing material, is first distributed to an influencer then passed to the network according to . In such a two-step scheme, maximizing the log-likelihood means to maximize the influence coverage. The maximum likelihood solution means an optimal scheme to allocate the information source. It tells that some influencers with large weights are preferred among the others in distributing information. By eq. (3), means that given an influenced node , how likely such influence comes from each influencer. Therefore, means the percentage of actual influenced nodes in by each influencer. Minimizing in eq. (3) means that the random influencer should be placed according to its effective influence.

**Diversifying** The objective of diversification corresponds to sparsity of as discussed in section 2. Making sparse, i.e. making certain influencers deactivated, helps to save resources in real world applications. For example, only a limited number of marketing personnel have to be deployed for broadcasting a piece of information.

In the following, we study several simple cases of the proposed optimization, so that one can better understand the result ranking.

**Proposition 6** Consider the influencers where is the set of predecessors. Denote the optimal solution of eq. (3) on the whole graph as . Denote the optimal solution on the sub-graph induced by as . Then for all , in .

The condition in proposition 6 means that can be partitioned into sub-graphs so that any two sub-graphs do not share a common influencer. Proposition 6 says that the distribution of among the sub-graphs is proportional to the size of the target population. A random influencer is more likely in populated regions. In a sub-graph, if the number of influencers is large but the target population is small, there will be more competitions among the influencers, in the sense that only a small percentage of influencers will be activated. The optimal can be obtained by individually solving eq. (3) on the sub-graphs and assembling them based on proposition 6.

**Proposition 7** for all , where means the isolated predecessors of having only as their successor.

Proposition 7 gives an upper bound and a lower bound of . Consider that and should be similar, this gives an estimated range on the resulting . An influencer with at least one isolated predecessor cannot be deactivated, because its isolated predecessor(s) can only be influenced by it. An influencer with many isolated predecessors is likely to have a large value of and a high rank.

**3.1. Implementation**

We implemented an optimizer for eq. (3), focusing on the case with very efficient optimization. This is because has the simplest expression and gives the most sparse solution, which is preferred in our social analysis. We call this ranking method “diversify”.

22
The optimizer is based on the natural gradient (Amari, 1998). Recall from section 2 that the natural gradient of a function is a vector field that is invariant to the choice of the coordinate system. We choose the spherical coordinates \( \beta = (\beta_0, \beta_1, \ldots, \beta_m) \), so that \( \forall j = 0, \ldots, m, \beta_j = \sqrt{\eta_j} \). Lebanon (2003) proved that FIM is “equivalent” to the embedded Euclidean geometry on the hyper-sphere \( \{ \beta \} \). We further have the following propositions.

**Proposition 8** For any smooth function \( f \) on \( S^m \),

\[
\nabla f = \frac{1}{4}(I - \beta \beta^T) \frac{\partial f}{\partial \beta} \cdot \partial \beta = \frac{1}{2} \sum_{j=0}^{m} \beta_j \left( \frac{\partial f}{\partial \eta_j} - \sum_{j=0}^{m} \eta_j \frac{\partial f}{\partial \eta_j} \right) \frac{\partial \beta}{\partial \eta_j}, \tag{6}
\]

where \( I \) is the identity matrix.

**Proposition 9** \( \forall j, \frac{\partial E_0}{\partial \eta_j} = -\sum_{i:i\rightarrow j} \eta_i^j (1 + \log \eta_j - \sum_{j=0}^{m} \eta_j \log \eta_j) / (n \eta_j) \).

Equation (6) is exactly the projected gradient on the hypersphere \( \{ \beta \} \) up to constant scaling. Proposition 8 gives an easy way to apply and understand natural gradient. To optimize any cost function \( f(\eta_0, \ldots, \eta_m) \) on \( S^m \), one can regard \( \eta_0, \ldots, \eta_m \) as independent variables, compute \( \frac{\partial f}{\partial \eta_j}, \forall j = 0, \ldots, m \), and then compute the natural gradient by proposition 8.

First, \( \beta^0 \) is randomly initialized, so that the corresponding \( \eta^0 \) is roughly uniform for all influencers. Then we update \( \beta^{t+1} (t = 0, 1, \ldots) \) until convergence following the rule

\[
\beta^{t+1} \leftarrow \beta^t - \gamma \nabla E_0 \frac{\beta^t}{||\beta^t - \gamma \nabla E_0||},
\]

where \( \gamma > 0 \) is a small learning rate, \( || \cdot || \) is 2-norm, and \( \nabla E_0 \) is given by propositions 8 and 9. An intuitive explanation of the learning process is in subsection 4.2.

Despite that the proposed optimization is non-convex and hence does not guarantee a global optimum solution, we find that its convergence is quite fast. To get a rough idea, fig. 2 shows the evolution of \( E_0 \) in the first 100 iterations on a DBLP collaboration network and an autonomous system network. In both cases, the convergence is reached in \(~30\) iterations. Such fast convergence guarantees scalability. By our C++ implementation, to compute \( \beta \) on a graph of \(~1.5\) million nodes only costs minutes on a normal PC.

![Figure 2: \( E_0 \) against the number of iterations](image)
4. Discussion on Related Works

4.1. Bayesian methods

From a Bayesian perspective (Bishop, 1995), the i.i.d. observations \{X^i\} induce on \(S^m\) a posterior distribution \(p(\eta | \{X^i\}) \propto \prod_i \{\sum_Y p(X^i | Y)p(Y | \eta)\} p(\eta)\). It is natural and interesting to view such a Bayesian inference on \(S^m\) as a manifold with boundary.

Consider the scenario to find an optimal \(\eta^* \in S^m\). The log-likelihood \(L = \sum_i \log(\sum_Y p(X^i | Y)p(Y | \eta))\), or a learning cost function derived from \(L\), is in general smooth on a feasible region \(\mathcal{F} \subset S^m\), meaning that deactivating some bits in \(Y\) could bring down the value of \(L\) but does not create singularities. The learning dynamics near \(\partial S^m\) largely depend on the prior distribution \(p(\eta)\). Traditional maximum likelihood learning uses flat priors, i.e. \(p(\eta) \propto 1\), resulting in a learner who is unaware of \(\partial S^m\). This is shown by the smoothness of \(E_1\) in theorem 3. Jeffrey’s prior (1946) is proportional to the Riemannian volume element (Lee, 2012) such that \(p_J(\eta) \propto |g(\eta)|^{1/2}\). It is non-informative, treating different points on \(S^m\) intrinsically equally. We give without proof that \(p_J(\eta) \propto \prod_{j=0}^m \eta_j^{-1/2}\). This leads to an inward flow that is similar to theorem 4 due to the non-smoothness of \(\log t\) at \(t = 0\). Such a similarity could partially justify the setting \(\tau < 1\) used in this paper.

Because \(p_J(\eta) \to \infty\) as \(\eta \to \partial S^m\), \(p_J(\eta)\), as well as \(p(\eta | \{X^i\})\), is not continuous on \(S^m\). It is easy to see that the prior used in this paper is \(p_D(\eta) \propto \prod_{j=0}^m \eta_j^{(1-\tau)\eta_j}\). Similar to \(p_J(\eta)\), it has a concave shape on \(S^m\). The difference is that it dampens \(p_J(\eta)\) near \(\partial S^m\) with a finite value on \(\partial S^m\), meaning that sparsity instead of small values of \(\eta_j\) is preferred. It yields a continuous \(p(\eta | \{X^i\})\) on \(S^m\), which is elegant in theory and establishes a global Bayesian view on \(S^m\).

The model \(S^m\) can be assessed by evaluating \(p(\{X^i\}) = \int_{\eta \in S^m} p(\{X^i\} | \eta)p(\eta)d\eta\). If \(p_J(\eta)\) is used, the boundary regions occupy a large percentage of the total volume. For example, the volume of the region \(\{\eta \in S^m : \exists j, \eta_j < 0.05 \text{ or } \eta_j > 0.95\}\) is at least \((1-0.9^m)\) times the total volume of \(S^m\). \(p_J(\eta)\) as a non-informative prior emphasizes too much on such regions. As a result, the model assessment is largely based on such \(\eta\) with many small non-zero values of \(\eta_j\)’s. This can be understood as a curse of dimensionality (Bellman, 1957) on the parameter manifold. \(p_D(\eta)\), as a weakly informative prior, puts focus on the center of \(S^m\) with sufficient large \(\eta_j\)’s. It could favor a simple model \(S^m\) over a complex model \(S^{m+1}\), if the maximum likelihood solution on \(S^{m+1}\) is near \(\partial S^{m+1}\), while the maximum likelihood solution on \(S^m\) is near the center. This essentially agrees with the idea of sparsity.

This work is connected to previous studies on singularities on statistical manifolds, where FIM is not well-defined (Amari et al., 2006; Cousseau et al., 2008; Park and Ozeki, 2009). Interestingly, we demonstrated that singularities can be helpful to learning, which is in contrast to the cases where singularities make learning difficult (Amari et al., 2006).

4.2. Diversified Ranking in Information Retrieval

Several proposals have been made for modeling and encouraging diversity in ranking. An extensive review of these works is proposed in (Raman et al., 2013). In summary, diversity is encouraged in ranked lists by modeling novelty and performing re-ranking (Carbonell and
Goldstein, 1998) following the Cascade Model of user behavior (Clarke et al., 2011). Query reformulation (Santos et al., 2010) also goes in the line of document re-ranking. Diversity may be associated with the inclusion of risk in the document ranking process. Risk may be viewed as ranking high novel but less-relevant documents (Wang and Zhu, 2009) or from the point of view of satisfying user intent (Agrawal et al., 2009). Alternatively, diversity may be embedded into the process of learning-to-rank by maximizing the expected user satisfaction over probable rankings (Radlinski et al., 2008; Slivkins et al., 2013).

**PageRank**

The widely-applied PageRank (Page et al., 1999) can be formulated by replacing the first two equations in eq. (3) with \( \tilde{\eta} = A^T \eta \), where \( A_{ij} = (1 - \nu)/n + \nu \delta_{ij}/\text{deg}(i) \), \( \delta_{ij} = 1 \) if \((i,j) \in E\) and \( \delta_{ij} = 0 \) if otherwise, and \( \nu = 0.85 \) is a damping parameter. In this case, the problem can be solved with fixed point iterations \( \eta \leftarrow \tilde{\eta} \). Comparatively, an updating scheme of \( \eta \) based on proposition 2 is given by

\[
\eta_j \leftarrow (1 - \gamma)\eta_j + \frac{\gamma}{n} \sum_{i=1}^{n} \left(1 + \theta_j - \theta^T \eta^i \right) \eta^i_j,
\]

(7)

where \( \gamma \) is a learning rate. They both can be understood as a voting process. In PageRank, each node \( j \) votes for its successors uniformly, weighted by the current \( \alpha_j \). The amount of votes received by each node \( j \) determines the new value of \( \alpha_j \). In eq. (7), the voting is neither uniform nor strictly positive. For each predecessors \( i \) of node \( j \), if \( \theta_j \) is smaller than the threshold \( (\theta^T \eta^i - 1) \), then node \( i \) casts a negative vote to node \( j \). In this way, \( i \) chooses strong candidates from its successors and penalizes weak candidates. A compact list of candidates can be elected. In the toy example in fig. 3, diversify only selects three influencers, while most nodes are deactivated. For example, node 6 in the graph receives zero weight, because its predecessor node 5 is already influenced by node 2.

---

5. This is only an intuitive view. The learning is in the \( \beta \)-coordinates as explained in section 3.1.
Table 1: SNAP datasets used in the experiments

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#nodes</th>
<th>#edges</th>
<th>Directed</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>p2p-Gnutella04</td>
<td>10,876</td>
<td>39,994</td>
<td>Yes</td>
<td>Gnutella peer to peer network</td>
</tr>
<tr>
<td>p2p-Gnutella05</td>
<td>8,846</td>
<td>31,839</td>
<td>Yes</td>
<td>Gnutella peer to peer network</td>
</tr>
<tr>
<td>p2p-Gnutella06</td>
<td>8,717</td>
<td>31,525</td>
<td>Yes</td>
<td>Gnutella peer to peer network</td>
</tr>
<tr>
<td>web-BerkStan</td>
<td>685,230</td>
<td>7,600,595</td>
<td>Yes</td>
<td>Web graph of Berkeley and Stanford</td>
</tr>
<tr>
<td>soc-Pokec</td>
<td>1,632,803</td>
<td>30,622,564</td>
<td>Yes</td>
<td>Pokec online social network</td>
</tr>
<tr>
<td>cit-Patents</td>
<td>3,774,768</td>
<td>16,518,948</td>
<td>Yes</td>
<td>Citation network among US Patents</td>
</tr>
<tr>
<td>com-DBLP</td>
<td>317,080</td>
<td>1,049,866</td>
<td>No</td>
<td>DBLP collaboration network</td>
</tr>
<tr>
<td>com-Amazon</td>
<td>334,863</td>
<td>925,872</td>
<td>No</td>
<td>Amazon product co-purchase network</td>
</tr>
<tr>
<td>com-Youtube</td>
<td>1,134,890</td>
<td>2,987,624</td>
<td>No</td>
<td>Youtube online social network</td>
</tr>
</tbody>
</table>

5. Experiments

In this section, we evaluate the proposed ranking algorithm on real data. We select several medium-to-large networks from Stanford large network dataset collection (SNAP). The datasets used cover various domains as shown in table 1.

5.1. Spread Information

We investigate how the ranking approaches can influence the network by spreading information. We select 5 different algorithms: Random (random selection), Indegree (ranking by indegree), Grasshopper (a graph-based ranking algorithm achieving both diversity and centrality (Zhu et al., 2007)), PageRank (Page et al., 1999) and Diversify. For each algorithm on each dataset, we extract the top-ranked nodes (seeds) and count the number of nodes that can be influenced by them, that is, the number of nodes that link to these seeds. This performance measurement makes sense in numerous applications, such as finding valuable nodes for content distribution networks (CDN), or finding impactive patents in patent-citation networks. Simply picking the mostly-linked nodes is likely to fail in such tasks, because a pair of well-connected nodes often have high-overlap in their influence.

Figure 4 shows the number of influenced nodes against the number of seeds. It is clear that the proposed algorithm most effectively covers the network, followed by Grasshopper, PageRank and Indegree. Random performs the worst as expected. Grasshopper selects nodes in a greedy manner. To select each node requires a large matrix inversion. It fails to operate in reasonable time on large datasets with millions of nodes, and thus no corresponding result is shown.

The good performance of Diversify as compared to Grasshopper that is designed for a similar purpose is explained as follows. Diversify is capable of global coordination: an influencer with large indegree but bad cooperation can be kicked out. Grasshopper is a greedy algorithm. If a bad influencer is already selected, there is no way to reverse it. Diversify performs better in network coverage and computation efficiency.

5.2. Community Coverage in Social Networks

Consider social networks where the users form communities. In many applications, the goal of ranking is to identify representative individuals (Chen et al., 2009; Sun et al., 2013), who are both well-connected individually, and involve in diverse communities. For example, in scientific collaboration networks, it is a common task to select researchers in distinct sub-areas for organizing events. We use datasets with ground-truth communities and keep these ground-truth from the ranking algorithms. We evaluate the ranking results by counting the number of communities covered by the top-ranked nodes.

We select 3 social networks: 1) DBLP computer science collaboration network, where the links represent co-authorships between authors, and the communities are defined by publication venue, etc. 2) Amazon product network, where the links represent frequently-occurred co-purchases between products, and the communities are defined by product categories.
3) Youtube social network, where the links are user friendships, and the communities are user-created groups on this platform. For each dataset, 5000 high quality ground-truth communities are known ahead (Yang and Leskovec, 2012). Unfortunately, we did not find any publicly-available directed graph with ground-truth community information. The proposed approach can handle these undirected graphs by replacing each edge with two opposite-directed links. We remove Grasshopper from the comparison, because it does not scale well on large datasets as shown in section 5.1. We compare another technique, Degree-discount (Chen et al., 2009), a simple heuristic for undirected graphs, which can effectively select nodes to maximize their influence spread.

Figure 5 shows the number of communities covered by the top-k-ranked nodes against k, where k ranges from 100 to 10,000. We assume that for even larger values of k, the community coverage is less interesting, because selecting more seeds usually implies costing more resources. On small values of k, all methods perform similarly. The difference shows up as k increases. We see that in general Diversify is among the top methods that cover the largest number of communities. Degree-discount also achieves good performance as compared to PageRank and Degree on the first two datasets. The good performance of Random on the Amazon dataset is because the product-categories are small-size communities which are likely to be disjoint. A large co-purchase number does not guarantee high coverage of such communities. This is different from social networks, where popular individuals tend to belong to more communities. However, the seed quality by Random is expected to be lower than Diversify, because Diversify considers the link structure and selects nodes with high degrees. On Youtube, the community distribution is much more sparse. Around 90% individuals are not signed up in any communities. If k is below 5,000, counting the most active individuals by Indegree effectively covers different communities. On the range from 5,000 to 10,000, Diversify still performs best (note that x-axis is log-scale).

5.3. Graph-based Movie Ranking

To apply the proposed method to a real-world ranking scenario, we select the MovieLens dataset7 consisting of ~10 million 5-star-ratings from ~72,000 users to ~10,000 movies. We check whether each user gives at least 4.5 stars to each movie, resulting in a directed user-rate-movie graph with 78,377 nodes and 2,129,834 edges. We also check whether each pair of movies are simultaneously rated higher than 4.5 by at least 10 users. The largest connected component gives an undirected movie-co-like graph with 5,316 nodes and 1,594,531 edges.

Table 2 shows the top-ranked movies based on the co-like graph 8. The top-15 by PageRank concentrated on movies in the 1990s. There are two episodes of “Star Wars”, which are similar in contents. Diversify presents a wider range in the sense of release-time. It could therefore satisfy more users. It discovers non-English movies like “Wooden Man’s Bridge”. An interesting observation is that it selects more old classical movies like “Citizen Kane”. Such an observation is consistent when we vary the settings such as the threshold of simultaneous co-likes (which is set to 10 in this experiment) used to construct the co-like graph. Table 3 shows the coverage of the top-ranked movies. Movie coverage is based on the number of movies directly linked to the top list in the movie-co-like graph.

---
7. The 10M dataset at http://grouplens.org/datasets/movielens/ is used.
Table 2: The top-15 movies on the MovieLens dataset based on co-like relationships. The unique movies discovered by Diversify are displayed in bold.

<table>
<thead>
<tr>
<th>Rank</th>
<th>PageRank</th>
<th>Diversify</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Godfather (1972)</td>
<td>Godfather (1972)</td>
</tr>
<tr>
<td>10</td>
<td>Raiders of the Lost Ark (1981)</td>
<td>Citizen Kane (1941)</td>
</tr>
<tr>
<td>14</td>
<td>Usual Suspects (1995)</td>
<td>Dr. Strangelove (1964)</td>
</tr>
</tbody>
</table>

Table 3: Movie coverage among 5,316 movies and user coverage among 68,860 users by the top-ranked movies. The “Sparsity” column shows the percentage of movies weighted greater than $10^{-7}$ based on the rankings in the co-like graph.

<table>
<thead>
<tr>
<th></th>
<th>Movie Coverage</th>
<th>User Coverage</th>
<th>Sparsity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>top-10</td>
<td>top-100</td>
<td>top-10</td>
</tr>
<tr>
<td>PageRank</td>
<td>5055</td>
<td>5236</td>
<td>50497</td>
</tr>
<tr>
<td>Diversify</td>
<td>5058</td>
<td>5301</td>
<td>52042</td>
</tr>
</tbody>
</table>

User coverage is based on the number of users linked to the top list in the user-rate-movie graph. In both cases, Diversity is able to cover more movies or users. Unlike PageRank, its ranking is sparse, meaning that the majority movies receive a score of zero. This could be useful in scenarios such as purchasing a small number of representative movies.

6. Conclusion

We present an information geometric analysis on the sparsity on $S^m$. The discovery of the inward flow near $\partial S^m$ helps to understand the learning dynamics and the model variation. We advocate a weakly informative prior derived from the negative entropy function. It is continuous on $S^m$ and is meaningful in Bayesian inference and model selection. We apply the proposed sparsity technique on graph-based ranking problems to enforce diversity. It scales to social graphs with tens of millions of nodes. In our experiments, the proposed method most effectively covers social networks among several commonly-used techniques.
Acknowledgments

This work is supported by the European COST Action on Multilingual and Multifaceted Interactive Information Access (MUMIA) via the Swiss State Secretariat for Education and Research (SER grant C11.0043).

References


Sparsity on Statistical Simplexes


