A Dynamical System on Bipartite Graphs

Kishore Papineni  
Google Research  
111 8th Ave  
New York, NY 10011, USA  

Pratik Worah  
Google Research  
111 8th Ave  
New York, NY 10011, USA

Abstract

This paper poses a non-linear dynamical system on bipartite graphs and shows its stability under certain conditions. The bipartite graph is composed of a set of words and a set of sentences or documents (containers) with edges between words and sentences representing containment. The dynamical system changes the weights on the nodes of the graph in each time step. The underlying weight transformation is non-linear, motivated by information gain in a document retrieval setting. Stability analysis of this problem is therefore different than that of PageRank-like algorithms. We show convergence using methods from Lyapunov theory and also provide some examples of how the algorithm performs when ranking keywords and sentences in a set of documents.

1. Introduction

1.1 Motivation

Suppose we are to summarize a corpus of documents that do not have explicit hyperlinks among them. Arguably, a good summary would contain important concepts from the corpus and arise from important sentences in these documents. Given a collection of words and sentences from a corpus of related documents, our goal then is to associate a weight with each sentence and each word (or phrase) which can be interpreted as their importance in the corpus.

From the viewpoint of natural language understanding, words that occur too frequently are usually not important. For example, words such as ‘the’ are usually not important and typically will not be the keywords of a document. This is the informal intuition behind Inverse Document Frequency (IDF) Sparck-Jones (1973). However, IDF is the highest for rarest words; but just as very frequent words in a document are unimportant, very rare words are also not expected to be keywords of a document. Therefore, neither too frequent nor very rare words are going to be keywords in a document, and the sweet-spot lies somewhere in the middle.

The above informal intuition was first described in Luhn (1958), and was captured in Papineni (2001) in a mathematical setting, which defined the information gain of a word as the Kullback-Leibler divergence of the word’s relatively frequency of occurrence in
a document. In other words, the information gain from a given word is

\[ g_w := f_w \cdot (f_w - 1 - \ln(f_w)), \]

where \( f_w \) denotes the fraction of sentences containing the word \( w \). Note that the gain, as a function of \( f \), is a positive and inverted U-shaped function in \([0, 1]\) (the range of \( f \)), with a maximum near 0.2. Of course, there is nothing magical about the constant 0.2 i.e., ideal keywords are not necessarily those that occur with a frequency of 20\%. This particular fraction is merely an artifact of using equal weights on all sentences to measure the Kullback-Leibler divergence. Therefore, assigning the correct weights to sentences is important before we can identify keywords. In particular, long sentences with many keywords should intuitively be more important from a natural language understanding and summarization point of view than short sentences which contain mostly filler words. Hence, we are now facing a circular problem: We need to identify important sentences before we can identify important words, but to identify important sentences we need to identify important words.

The circular nature of our problem suggests the following fixed-point type algorithm to compute a weighted version of information gain: Assign a (reasonable) initial weight to all words and sentences. For each word \( w \) and each sentence \( s \), compute \( \text{sum}_w \) and \( \text{sum}_s \) – the sum of weights of sentences that contain \( w \) and the sum of weights of words in \( s \). Update the weight of each \( w \) and \( s \) as follows:

\[
\text{Weight}_w \leftarrow \text{sum}_w \cdot (\text{sum}_w - 1 - \ln(\text{sum}_w)),
\]

\[
\text{Weight}_s \leftarrow \text{sum}_s \cdot (\text{sum}_s - 1 - \ln(\text{sum}_s)).
\]

Repeat the updates until the weights of all vertices do not change significantly between two iterations.

Mathematically, it is not clear under what conditions i.e., for what kinds of corpus and what initial choice of weights, does the above algorithm converge to a “optimal” weight distribution. Intuitively, it is not clear how the important sentences and keywords generated by using such a weighted information gain qualitatively compare with some other conventional approach, or even in some absolute sense.

This paper applies basic linear algebra and Lyapunov theory to prove convergence of the above iterated update algorithm when the underlying word-sentence graph obeys certain eigenvalue constraints. Moreover, we have also implemented the above fixed-point algorithm and ran it on a small corpus of documents to see if there was a natural qualitative difference between the sentences produced with weights and using a conventional method\(^1\). Below are a couple of examples\(^2\).

The top sentence produced using the weighted reflection algorithm using the top twenty webpages in a web search for “monarch butterfly parasites”.

**Protozoan parasites such as**

**Ophryocystis elektroscirrha and a**

**microsporidian Nosema species**

---

1. Comparison with unweighted information gain is not very interesting as it merely picks the words and sentences around the 20\% frequency range as explained above.
2. The maximum number of characters in any output sentence was capped at 140.
have also been identified
in wild and captive monarchs
(McLaughlin and Myers 1970,
Leong et al. 1992; 1997,
Altizer and Oberhauser 1999,
O. Taylor, personal communication).

The top sentence produced using the Maximal Marginal Relevance (MMR) algorithm Carbonell and Goldstein (1998) using the same input as above is given below.

Monarchs have many natural enemies -
predators, parasitoids, and parasites
can harm monarch eggs, larvae, pupae,
and adults.

The sentences produced using weighted information gain are typically longer, containing diverse concepts. The remaining sections of the paper provide more details about our algorithm and its properties. The last section discusses a few more examples from our experiments.

Finally, we observe that there is no reason why the algorithm can’t be applied to more general settings. For instance, instead of a collection of words and sentences, we can apply it to any container which is a collection of tokens – a corpus of images and the entities within those images; a corpus of videos and the entities within those videos, and so on.

1.2 Related Work

The fixed-point approach described in the introduction implicitly assumes a graph structure between sentences and words in a document. In the underlying graph, every word and sentence corresponds to a vertex, and a word is linked to a sentence by an undirected edge, if it occurs in the sentence. There is already a significant amount of work which applies fixed-point algorithms over an underlying graph structure to compute a rank or score which has some relevance in the real world. For example, the initial papers Page et al. (1998); Kleinberg (1999) in the areas of citation analysis, social networks and analyzing the link structure of the world-wide web, use similar ideas. More importantly, in the area of lexical analysis and ranking of documents there has been significant work using such graph based models, see for example Mihalcea (2004); Mihalcea and Tarau (2004) and the book Mihalcea and Radev (2011).

In Page et al. (1998), the authors work on a graph where the vertices are webpages and the edges reflect the link structure of the web. If \( \delta_+(v) \) and \( \delta_-(v) \) denote the out-degree and in-degree neighbours of a vertex \( v \), then the score of \( v \) (on which its Page-rank is based) is defined as:

\[
S(v) := (1 - c) + c \cdot \sum_{u \in \delta_-(v)} \frac{S(u)}{\delta_+(u)},
\]

where the constant \( c \) is around 0.8. The idea is to iterate until the score stabilizes and then use the scores to rank web-pages. Note that the initial scores do not matter too much as the scores converge rapidly Page et al. (1998).
Kleinberg (1999) introduced a similar algorithm (HITS), which ranked webpages by implicitly ascribing a bipartite structure to the graph – vertices correspond to “authority” pages (corresponding to a large $\delta_-$) or hub pages (corresponding to a large $\delta_+$). Therefore, we have two scores for each vertex: an authority score and a hub score – both calculated as a weighted linear sum over their neighbours.

More relevant to us, Mihalcea and Tarau (2004) introduced TextRank an unsupervised procedure to extract and rank keywords and sentences, or more generally text units, from a lexical corpus. Their approach is similar to Page’s approach, but now the graph vertices correspond to text units (for example, sentences or phrases) and edges reflect some semantic or syntactic connection between text units. While the algorithm and score used is similar to that in Page et al. (1998), one important difference was that the edges can be weighed and so their counterpart of Equation 1 would be:

$$ S(v) := (1 - c) + c \cdot \sum_{u \in \delta_-(v)} \frac{W_{uv} \cdot S(u)}{\sum_{w \in \delta_+(u)} W_{uw}}, $$

where $W$ is the weight matrix of the underlying graph and the constant $c$ may be chosen depending upon the exact scenario at hand.

While the viewpoints in each case may be different, the relevant theme, at least for us, remains the same i.e., there is an iterated local computation on a graph that leads to a score for each vertex, and this score is interesting from the perspective of some real world problem. However, in all the above cases the score is computed using a linear function, as in Equations 1 and 2, for example. Linearity together with high connectivity in the graph structure leads to rapid mixing, thus convergence is ensured and is typically very fast. This linearity is in contrast to our situation, where we iteratively update based on a Kullback-Leibler divergence type of function – a non-linear function, and so convergence does not follow from earlier ideas. Therefore, we go back to first principles and analyze convergence of the underlying dynamical system using Lyapunov theory.

2. Preliminaries

Before we dive into our formal results, we present some basic and essential definitions related to graphs (see for example, West (2000)) and dynamical systems (see for example, Khalil (2002) or Frazzoli and Dahleh (2011)).

A graph $G$ consists of a set of vertices $V$, and a set of edges $E \subseteq V \times V$. In this paper we are mainly concerned with undirected graphs i.e., the relation defining $E$ is symmetric. The neighbors of a vertex $v$ is the set of vertices $N(v) \subseteq V$ such that each member of $N(v)$ is connected to $v$ by some edge $e \in E$. Additionally, the edges in the graph may be weighted according to some weight function $W : E \rightarrow \mathbb{R}$. The adjacency matrix of $G$ is a (symmetric) matrix representation of $W$, where an edge $(u, v) \in E$ is assigned a weight $W_{uv}$, all other entries are 0. Note that every symmetric adjacency matrix corresponds naturally to a graph.

**Definition 1** A bipartite graph (also bigraph), is a graph where the vertex set can be partitioned into two disjoint components such that there are no edges joining two vertices in the same partition.
Definition 2 The square of a graph $G$ is the graph whose adjacency matrix is the square of the adjacency matrix of $G$.

Suppose $z(t) \in \mathbb{R}^n$ describes the state of a dynamical system, which is just an ordinary differential equation of the form $\frac{d}{dt}z(t) = f(z(t))$, $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ being a continuous function.

Definition 3 A dynamical system is said to be Lyapunov stable around its fixed point, assumed to be the origin, if for every $\epsilon > 0$, there exists $\delta > 0$ such that if $\|z(0)\| \leq \delta$ then $\|z(t)\| \leq \epsilon$.

Definition 4 A dynamical system is said to be asymptotically stable around its fixed point, assumed to be the origin, if it is Lyapunov stable and it satisfies: $\exists \delta > 0$, such that if $\|z(0)\| \leq \delta$ then $\lim_{t \rightarrow \infty} \|z(0) - z(t)\| = 0$.

We will introduce and explain further background details as needed and the references mentioned above should provide any background definitions, if necessary.

3. Results

In this section, we recap the mathematical model for our problem and then provide a proof of Theorem 5 and 12, our main mathematical results. We provide experimental verification in the next section.

We are given an undirected bipartite graph $G \equiv (X, Y, W)$, where $X$ and $Y$ denote the vertex sets and $W$ denotes the edge weights i.e., $W_{uv} \in \mathbb{R}$ is the weight of the edge connecting vertices $u$ and $v$ in $G$. With $|X| = n$ and $|Y| = m$, $W$ is a $(n + m) \times (n + m)$ block offdiagonal symmetric matrix. Let $x_u(t) \in \mathbb{R}^n$ and $y_v(t) \in \mathbb{R}^m$, where $u \in X$ and $v \in Y$, denote the values of the weight distribution at time $t$ on the vertices $u$ and $v$ respectively. Let

$$z(t) := \begin{pmatrix} x(t) \\ y(t) \end{pmatrix},$$

denote the column vector of all weights, which is the state of the dynamical system that we will induce on the graph.

Following Papineni (2001), we define the time-invariant gain function $g : \mathbb{R} \rightarrow \mathbb{R}$ as follows:

$$g(x) := x \cdot (x - 1 - \ln|x|), x \in \mathbb{R},$$

define the vector version $\bar{g} : \mathbb{R}^{n+m} \rightarrow \mathbb{R}$, as the component-wise application of $g$ to each element in the argument.

At time $t + \Delta t$, we update the weight vectors $x$ and $y$ based on the following update rule:

$$z(t + \Delta t) = \bar{g}(Wz(t)).$$

Note that edge weights are constant, only the vertex weights change over time. We thus have a nonlinear time-invariant dynamical system on the graph. We make a mild

---

3. In our motivating application, $W$ is the adjacency matrix with $w_{uv} \in \{0, 1\}$. 

---
assumption that $W$ is full-rank. In our motivating example, this assumption is satisfied when no two tokens appear in exactly the same set of containers. If they do, we can replace by a joint token. When $W$ is non-binary, we can perturb $W$ to make it full-rank. This concludes the description of our underlying model. We can state our main theorem as follows.

**Theorem 5** Assume that there exists $z_e := (x_e, y_e) \in \mathbb{R}^{n+m}$, a positive solution to the system $\bar{g}(Wz_e) = z_e$. Suppose that $\lambda$ denotes the maximum eigenvalue of the graph $G^2$ obtained by squaring $G$ i.e., squaring the adjacency matrix of $G$, and then multiplying the weight of each edge $uv \in G^2$ by $g'(z_e(u)) \cdot g'(z_e(v))$. The algorithm for updating vertex weights given above converges to some stable point $z_e$ if $\lambda < 1$, and the initial value $z_0$ lies in the basin of attraction of $z_e$.

**Remark 6** It is worth noting that the rate at which $\Delta t$ goes to 0 does not effect the result i.e., if $\Delta t$ were replaced by $2\Delta t$, the condition on eigenvalues does not need to be scaled.

**Remark 7** Existence of $z_e$: If $W$ is full-rank then there exists a solution $z$ to $Wz = x$, where $x \in (0, 1]^{n+m}$. If the weights of $W$ are positive and small enough, then $z \in \mathbb{R}^{n+m}_+$ exists. As $x \to \infty$ in each component so does $z$ (linearly in $x$ since $z = W^{-1}x$), and also $\bar{g}(x)$ (but quadratically in $x$). Therefore, by continuity, $z$ must intersect $\bar{g}(x)$ at some value of $x$ – that’s a positive solution to $\bar{g}(x) = z$.

**Remark 8** Squaring the bigraph $G$ has the effect of “creating” two graphs, one on each set $X$ and $Y$, where the weights on the edges in $G^2$ are given by the (weighted) number of length two walks in $G$. It is worth noting that it’s the interconnection between the documents and interconnections between the words that separately determine the convergence criterion.

We first sketch the main ideas behind the proof in the scalar case before presenting the actual proof. Consider the following scalar ODE:

$$\frac{d}{dt} x(t) = \bar{g}(x) := \bar{g}(x) - x(t) = x(t) \cdot (x(t) - 1 - \ln x(t)) - x(t). \quad (6)$$

The rate of change in weight $x(t)$ in time $\Delta t$ is equal to the change in $x(t)$ in one update step, and as $\Delta t \to 0$, we get the ODE above. The ODE has two fixed points: (1) $x = 0.158...$, and (2) $x = 3.146...$, with $\bar{g}$ as shown in Figure 1.

Our goal is to analyze whether either of the fixed points is locally asymptotically stable i.e., if the weights start out close enough to the fixed point then as $t \to \infty$ the weights will remain close to that fixed point. In order to analyze each of the above two fixed points for asymptotic stability it is best to translate the system so that the fixed point is at the origin. Therefore, we translate the graph to the left to set $x_e := 0.158...$ as the origin and get the ODE:

$$\frac{d}{dt} x(t) = \bar{g}^{(x_e)}(x) := x(t + x_e) \cdot (x(t + x_e) - 1 - \ln(x(t) + x_e)) - x(t) - x_e. \quad (7)$$

Next, basic Lyapunov theory says that if there exists a locally positive definite function $V(x)$, such that $V(0) = 0$ and $V(x)$ is locally negative definite then the origin is a locally
A Dynamical system on Bipartite Graphs

\[ \begin{align*}
\text{Figure 1: } y &= x(x - 1 - \ln x) - x \\
\end{align*} \]

asymptotically stable fixed point. Let

\[ V(x) := -\int_0^x \hat{g}(xe)(x) \, dx, \quad (8) \]

then it is easy to see that \( V(0) = 0 \) and \( \dot{V}(x) = -\hat{g}(xe)(x) \cdot \hat{g}(xe)(x) \), so that it is locally negative definite for \( x \leq 3.146 - 0.158 \simeq 3.0 \). Hence, by Lyapunov’s criteria such a one dimensional update rule for \( x(t) \) would lead to convergence with \( x^* \simeq 0.158 \ldots \), as long as the initial value \( x(0) \) was close to \( x^* \).

The analysis of such a dynamical system in one dimension is simple because the system is memoryless so it can do one of two things – converge to a point or run away to infinity. However, in two and higher dimensions, \( n + m \) in our case, the problem is more difficult to analyze, because the solution is not confined to just the axis, it may escape to infinity from in-between. Still the foundation of our analysis remains the same – basic Lyapunov theory. We need the following indirect version of Lyapunov’s stability criterion for higher dimensions:

**Theorem 9 (Khalil (2002))** Lyapunov stability criterion: Let \( x = 0 \) be an equilibrium point for \( \dot{x} = f(x) \), where \( f : D \to \mathbb{R}^n \) is continuous and differentiable and \( D \) is a neighborhood of the origin. Let \( J \equiv \frac{\partial}{\partial x} f(x)|_{x=0} \) denote the Jacobian matrix, then

- The origin is asymptotically stable if the real part of all eigenvalues of \( J \) are negative.
- The origin is unstable if the real part for at least one of the eigenvalues is positive.

Consider the following \( n + m \) dimensional ODE:

\[ \frac{d}{dt} z(t) = \hat{g}(z(t)) := \hat{g}(Wz(t)) - z(t). \quad (9) \]

The above dynamical system captures our update rule. Note that any \( z_e \in \mathbb{R}^{n+m} \) that satisfies \( \hat{g}(Wz_e) = z_e \) will be a candidate for a stable fixed point. Since in one dimension
we have two fixed points, we may potentially have many candidates for stable fixed points. In the analysis below, we assume the existence of at least one fixed point $z_e$.

Our next task is to characterize the eigenvalues of the Jacobian at $z_e$. The $uv$th entry of the Jacobian matrix at $z_e$ is given by:

$$\frac{\partial}{\partial z_v} \left( g \left( \sum_{w \in N(u)} W_{uw} z_w \right) - z_u \right) \Bigg|_{z_e},$$

where $N(u)$ denotes the set of neighbours of vertex $u$. By the chain rule, it equals:

$$\left( \frac{\partial}{\partial z_v} s(u) \right) \left( \frac{\partial}{\partial s} g(s) \right) \Bigg|_{z_e} - \delta_u(v),$$

where we have used $s(u) := \sum_{w \in N(u)} W_{uw} z_w$ and $\delta$ is the kronecker delta function. But, the last expression is equivalent to:

$$W_{uv} \cdot g'(s)|_{z_e} - \delta_u(v).$$

Therefore, the Jacobian about $z_e$ has the form:

$$J|_{z_e} = D \cdot \left( \begin{array}{cc} 0 & M^T \\ M & 0 \end{array} \right) - I,$$

where $M$ is the $n \times m$ block matrix which is the incidence matrix of the bipartite graph $G$, $I$ is the $n + m$ dimensional identity matrix, and $D$ is the diagonal matrix with its $u$th entry given by:

$$D_u = g'(\sum_{w \in N(u)} W_{uw} z_w(w)).$$

Let, $H := J + I$. If we can place an upper-bound on the real part of the eigenvalues of $H$ so that

$$\text{Re} \left( \lambda_{\text{max}}(H) \right) < 1$$

then the eigenvalues of $J$ are simply obtained by translation of the eigenvalues of $H$, so that $\text{Re}(\lambda_{\text{min}}(J)) < 0$ – the fixed point corresponding to $z_e$ is then stable. The condition which ensures Equation 15 will be our condition in Theorem 5.

Let us rewrite $D$ as:

$$D = \left( \begin{array}{cc} D_1 & 0 \\ 0 & D_2 \end{array} \right),$$

where $D_1$ and $D_2$ are $n \times n$ and $m \times m$ diagonal matrices, respectively.

**Proposition 10**

$$HH^T = \left( \begin{array}{cc} D_1 MM^T D_1 & 0 \\ 0 & D_2 M^T MD_2 \end{array} \right).$$
The eigenvalues of $HH^T$ are bounded by the maximum of the eigenvalues of $D_1MM^T D_1$ and $D_2M^TMD_2$. It follows from the definition of spectral norm that the absolute value of the eigenvalues of $H$ is upper-bounded by the square-root of the maximum eigenvalue of $HH^T$. Therefore, the condition that both matrices $D_1MM^T D_1$ and $D_2M^TMD_2$ have all eigenvalues upper-bounded by 1 is sufficient to ensure a stable fixed point.

However, the matrices $D_1MM^T D_1$ and $D_2M^TMD_2$ are simply the adjacency matrix of the graphs on the partite sets $X$ and $Y$ of $G$, obtained by squaring the weighted bipartite graph $G$ and reweighing every edge $uv$ in the squared graph by $g'(z_e(u))g'(z_e(v))$. Hence, if the eigenvalues of these graphs are all upper-bounded by 1, the statement of Theorem 5 follows.

### 3.1 An Example

As a concrete example, consider the case where the entries of $W$ are Gaussian random variables with zero mean and variance $\sigma$. This is a rather arbitrary assumption but this case is still worth pointing out. One mitigating factor is that we are merely interested in the fact that the spectrum is supported on a small interval and not on the distribution of eigenvalues, as $n, m \to \infty$.

**Remark 11** We note that in real world applications, negative weight between a word and a document could be used to signify that a document carries a diametrically polar opinion to the word, even if the word does not occur in the document.

Now, $HH^T$ is a positive semidefinite block matrix as before i.e.,

$$HH^T = \begin{pmatrix} D_1MM^T D_1 & 0 \\ 0 & D_2M^TMD_2 \end{pmatrix}. \tag{18}$$

It is known that the spectrum of $MM^T$ follows the Marcenko-Pastur distribution (Mehta, 2004) and it’s density has the form:

$$[1 - n/m]_+ \delta(x) + \frac{\sqrt{(\lambda_+ - \lambda_\pm)(\lambda_\pm - \lambda)}}{2\pi \lambda \frac{m}{M}} 1_{\lambda \in [\lambda_\pm, \lambda_\pm]}, \quad \lambda_\pm = \sigma^2 (1 \pm \sqrt{n/m})^2, \tag{19}$$

where $\delta$ is dirac delta function, and $1$ is the indicator function. Similarly one can obtain the limiting distribution for the spectrum of $M^T M$. Observe that if $\sigma \ll 1$, with $n$ and $m$ comparable to each other and $D$ is large enough, then the distribution is supported on an interval within $[0, 1)$. Therefore, such weight matrices will lead to a stable fixed point.

### 3.2 Basin of Attraction

The next task is to determine the starting values of $z$ for which we can be assured of convergence. In this case, the conceptually easiest solution is to construct a Lyapunov function, as in the one dimensional case above, and show that $V(z)$ is locally positive definite and $\dot{V}(z)$ is locally negative definite. However, things are more complicated in higher dimensions and such explicit constructions are not easy for non-linear dynamical systems. Therefore, the standard approach is to construct an ellipsoidal Lyapunov function for the linearized system, which is expected to be globally asymptotically stable, and then
show it is locally asymptotically stable in a large enough region. Typically, even this problem is difficult but the symmetries of the Jacobian matrix makes it simpler.

**Theorem 12** Given matrix $M$ which satisfies the conditions of Theorem 5, the radius of convergence about $z_e$ is non-decreasing in $\lambda$, where $\lambda$ is as in Theorem 5.

Consider a linear system $\dot{z} = Az$, it is asymptotically stable if the Jacobian $J$ has all eigenvalues with negative real parts. A more direct characterization is that if the linear system is asymptotically stable then for any positive definite matrix $Q$, there exists a positive definite matrix $P$ such that

$$A^T P + P A = - Q$$  \hspace{1cm} (20)

which follows from using $V(z) = z^T P z$ as the Lyapunov function Frazzoli and Dahleh (2011).

The same kind of analysis can be done for non-linear systems i.e., linearize the system about the fixed point (Taylor expansion with higher order terms dropped), compute a Lyapunov function which shows global asymptotic stability, then use it as a candidate for the non-linear system. $V$ already has all the properties we need except for one: $\dot{V}(z)$ may not be negative definite, but we can show it is locally negative definite and prove a lower bound on the convergence radius.

We know,

$$\frac{d}{dt} z(t) = J|_{z_e} z(t) + h(z(t)),$$  \hspace{1cm} (21)

where $h$ is what remains after substracting $J|_{z_e}$ from the RHS of Equation 9 i.e. $\bar{g}(z) - z$.

In our notation, $A = J|_{z_e}$, so that Equation 20 reads:

$$J^T P + PJ = -Q$$  \hspace{1cm} (22)

Choosing $Q = I$ gives the matrix equation:

$$J^T P + PJ = -I.$$  \hspace{1cm} (23)

Note that, the matrix $P$ in Equation 23 is assumed symmetric, so that the left and right eigenvectors coincide (after transposition). Multiplying the LHS and RHS of Equation 23 by $v_{\text{max}}$ and $v_{\text{max}}^T$, the unit eigenvector corresponding to the maximum eigenvalue $\lambda_{\text{max}}(P)$, we get,

$$\lambda_{\text{max}}(P) (v_{\text{max}}^T Jv_{\text{max}} + v_{\text{max}}^T J^T v_{\text{max}}) = -1.$$  \hspace{1cm} (24)

By the definition of spectral norm,

$$\max (|v_{\text{max}}^T Jv_{\text{max}}|, |v_{\text{max}}^T J^T v_{\text{max}}|) \leq \sqrt{\lambda_{\text{max}}(JJ^T)},$$  \hspace{1cm} (25)

and so,

$$\sqrt{\lambda_{\text{max}}(JJ^T)} \geq \frac{1}{2 \lambda_{\text{max}}(P)}.$$  \hspace{1cm} (26)
Now, it is also known (see (Khalil, 2002) or (Frazzoli and Dahleh, 2011)) that
\[
\dot{V}(z) = z^T(J^T P + PJ)z + 2z^TPh 
\]
\[
\leq -\|z\|^2 + 2\lambda_{\text{max}}(P)\|z\| \cdot \|h(z)\| \tag{27}
\]
\[
\leq - \left(1 - 2\lambda_{\text{max}}(P) \frac{\|h(z)\|}{\|z\|}\right) \|z\|^2, \tag{28}
\]
where \(\| \cdot \|\) denotes \(\ell_2\) norm. The RHS is negative as long as
\[
2\lambda_{\text{max}}(P) \frac{\|h(z)\|}{\|z\|} \leq 1, \tag{29}
\]
which is equivalent, by Equation 26, to
\[
\frac{\|h(z)\|}{\|z\|} \leq \frac{1}{2\lambda_{\text{max}}(P)} \tag{30}
\]
\[
\leq \sqrt{\lambda_{\text{max}}(JJ^T)}. \tag{31}
\]

As \(\lambda_{\text{max}}(JJ^T)\) increases \(\dot{V}\) is at least going to remain locally negative definite around \(z_e\). Recall from Proposition 10 and the discussion that followed that \(\lambda_{\text{max}}(JJ^T)\) is upper-bounded by the largest eigenvalue of the reweighed square of \(G\) i.e., \(\lambda \equiv \lambda_{\text{max}}(DW^2D^T)\). Therefore, as \(\lambda\) increases, the time derivative of the candidate Lyapunov function either remains negative or may become negative. Hence, the statement of Theorem 12 follows.

4. Experimental Results

The nature of our contributions is theoretical, although motivated by problems of relevance to natural language processing. We do not propose that the reflection algorithm by itself is a complete solution to a specific task like document summarization, but rather that the weights it induces on tokens and containers may be incorporated in applications like summarization. Human or automatic evaluation of a full document summarization system with and without the reflection algorithm is outside the scope of this paper. However, we present some qualitative comparisons with the MMR algorithm in the tables below.

Our comparison in the table simply displays the top ranked sentence for the set of top twenty documents obtained using google web-search\(^4\) for a given query, using the two algorithms, and lists them side-by-side. Below we present the set of top three keywords chosen by the reflection algorithm for the corresponding search phrases from the top-twenty documents.

In gathering the above data, each sentence is scanned for Freebase concepts and the identified span of an entity is replaced by its Freebase id. For example, ”There are three species of monarch butterflies” is rewritten as ”There are three /m/06zf0 of /m/01knvc”. We then represent each sentence as a weighted bag of tokens. In the MMR experiment, Freebase IDs were given a weight of 1.0 and all other words a weight of 0.1. In the reflection experiment, weights of tokens are the result of reflection algorithm.

\(^4\) Note that we exclude any wikipedia results in the top search results – it gives a reference set on which we can base our subjective judgement.
monarch butterfly parasites

Protozoan parasites such as Ophryocystis elektroscirrha and a microsporidian Nosema species have also been identified in wild and captive monarchs (McLaughlin and Myers 1970, Leong et al. 1992, 1997, Altizer and Olberhauer 1999, O. Taylor, personal communication).

monarch butterfly migration

While the practice of transferring monarchs from place to place is generally not condoned by scientists, some reciprocal transfers of tagged monarchs have demonstrated that monarchs from east of the Rocky Mountains will migrate south if transferred west, in the range of the western population (rather than SW).

monarch butterfly adult

The four stages of the monarch butterfly life cycle are the egg, the larva (caterpillar), the pupa (chrysalis), and the adult butterfly.

monarch butterfly climate

Aside from the ecological significance of these migrations—monarch butterflies are the only insects known to migrate to warmer climates more than 2,500 miles away—the butterflies’ five-month layover in Mexico before returning to the United States has become one of the region’s main tourist attractions and economic drivers.

monarch butterfly captive rearing

Moreover, these subtle variations appear to have biological significance; monarchs with darker shades of orange (approaching red) show higher flight ability in captive settings (13), and a recent study provided evidence that the degree of black pigment is related to migration distance in wild-caught monarchs (14).

<table>
<thead>
<tr>
<th>Query</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>monarch butterfly parasites</td>
<td>parasitism, infection, spore</td>
</tr>
<tr>
<td>monarch butterfly migration</td>
<td>animal migration, overwintering, butterfly</td>
</tr>
<tr>
<td>monarch butterfly adult</td>
<td>egg, caterpillar, monarch butterfly</td>
</tr>
<tr>
<td>monarch butterfly climate</td>
<td>climate change, species, prediction</td>
</tr>
<tr>
<td>monarch butterfly captive rearing</td>
<td>butterfly, larva, adult</td>
</tr>
</tbody>
</table>

Table 1: Some example sentences

5. Discussion

So far we have shown that convergence of our iterative updates method with a Kullback-Leibler gain function is determined by the maximum eigenvalue of the square of the (weighted) adjacency matrix of the document-word bipartite graph. Moreover, the radius of convergence is non-decreasing in the maximum eigenvalue of the same matrix. In our opinion this
A Dynamical system on Bipartite Graphs

is a surprisingly crisp result, since most results on such non-linear systems do not admit such a simple characterization. However, further investigation is needed to determine if intermediate normalization of weights would help increase the speed of convergence. It would also be interesting to bound the speed of convergence in terms of eigenvalues of the adjacency matrix (only), as has been possible for the case of linear update functions i.e., random walks on graphs. We suspect the speed remains exponential for expander graphs.

References


