# Nonnegative Matrix Factorization of Transition Matrices via Eigenvalue Optimization

**Braxton Osting**
Department of Mathematics
University of California, Los Angeles
Los Angeles, CA 90095
braxton@math.ucla.edu

**Chris D. White**
Department of Mathematics
University of Texas at Austin
Austin, TX 78712
cwhite@math.utexas.edu

## Abstract

We consider the nonnegative matrix factorization (NMF) approach to clustering where the matrix to be factorized is a transition matrix for a Markov chain. We prove the equivalence of this problem to an eigenvalue optimization problem and based on this equivalence, interpret clustering NMF as finding a $k$-partition of the data for which the stationary states of random walkers associated to each component are optimally closed. One novel feature of this interpretation is that it simultaneously outputs clusters as well as a "local ranking" of the data within each cluster, in the sense of PageRank. The local ranking provides label confidences and naturally identifies cluster representatives. A relaxed formulation is identified and a novel algorithm is proposed, which we show is strictly decreasing and converges in a finite number of iterations to a local minimum of the relaxed objective function. A semi-supervised version of the algorithm yields excellent results for the MNIST handwritten digit dataset. We conclude with an intriguing relationship to a reaction-diffusion system for antagonistically-interacting random walkers.

## 1 Introduction

Nonnegative matrix factorization (NMF) is the general problem of factorizing a matrix into nonnegative matrices which are constrained in various ways. Although NMF has a variety of applications, *e.g.*, variable selection [1, 2, 3] and low rank matrix approximation [4], in this paper we consider its application to clustering problems [5, 6, 7, 8]. Clustering NMF methods are closely related to spectral clustering techniques [5, 7] and exhibit state-of-the-art performance in terms of cluster purity on a wide collection of datasets [8].

Markov chains are a common tool used in machine learning; given a (not necessarily symmetric) transition matrix, $P$, constructed from data, properties of the associated random walk can be used to describe the data. Perhaps the most prominent example of this is PageRank [9], which utilizes the stationary distribution of the walk for statistical ranking. Another example is diffusion maps [10], which integrate the random walk at all time scales for the purpose of manifold learning.

In this paper, we consider the NMF approach to clustering, where the matrix to be factorized is a transition matrix of a Markov chain, $P$. Throughout, we assume $P \in \mathbb{R}^{n\times n}$ has nonnegative entries and that $P + P^t$ is irreducible. The NMF approach to cluster the dataset represented by $P$ solves

$$U^\star = \arg\min_{U\in\mathcal{A}} \|P - UU^T\|_F^2, \quad \text{where} \quad \mathcal{A} := \{U \in \mathbb{R}^{n\times k}\colon\ U_{ij} \geq 0,\ U^T U = \mathrm{Id}_k\}. \tag{1}$$

The constraint $U \in \mathcal{A}$ implies the columns of $U$ have disjoint support; the supports indicate the cluster labels. The objective in (1) is non-convex and consequently, most algorithms rely on heuristics, or complicated descent methods. Without loss of generality, we additionally assume the matrix $P$ to be symmetric.

In §2, we introduce the concept of *Perron-Dirichlet eigenvalues* which are defined on sub-matrices of $P$. We prove that (1) is equivalent to an eigenvalue optimization problem involving the Perron-Dirichlet eigenvalues. Based on this equivalence, we interpret NMF clustering as finding a $k$-partition of the data where random walkers associated to each partition component are *optimally closed*. One novel feature of this interpretation is that it simultaneously outputs clusters as well as a "local ranking" of the data within each cluster, in the sense of PageRank. This local ranking provides label confidences and naturally identifies cluster representatives, which can be used for automatic summaries, data compression, and visualization [11, 12].

In §3, we propose a novel algorithm for solving this equivalent eigenvalue optimization problem, based on an extension of recent work [13]. Moreover, our algorithm can naturally incorporate known information in the form of a semi-supervised extension. To our knowledge, this is the first NMF algorithm which can incorporate such information. We demonstrate the effectiveness of the proposed algorithm via a numerical experiment on the MNIST handwritten digit dataset. We conclude in §4 with an intriguing relationship between the eigenvalue optimization problem and a reaction-diffusion system for antagonistically-interacting random walkers on a graph.

The present work is most similar to [13], where the authors consider a graph partitioning objective based on Laplace-Dirichlet eigenvalues of the graph and prove an equivalence to NMF. The present work extends this connection to NMF and provides a probabilistic interpretation in terms of random walkers. We believe the geometric insight provided here could lead to an improved understanding of other algorithms.

## 2 Clustering NMF and an eigenvalue optimization problem

In this section, we show that Nonnegative Matrix Factorization (1) is equivalent to an eigenvalue optimization problem involving submatrices of $P$. The benefit of this equivalence is two-fold: we gain a novel interpretation of NMF in terms of random walkers, and in §3 we derive an easily implementable algorithm for approximately solving NMF.

Let $P \in \mathbb{R}^{n \times n}$ be a nonnegative, symmetric, irreducible matrix and $\Omega \subset [n]$. The *Perron-Dirichlet eigenvalue*, $\lambda_\Omega$, is the largest eigenvalue of the submatrix, $P_\Omega = [P_{ij}]_{i,j \in \Omega}$. These eigenvalues are also characterized by the variational formulation,

$$\lambda_\Omega = \max_{\substack{\psi \neq 0 \\ \psi|_{\Omega^c} = 0}} \frac{\langle P\psi, \psi \rangle}{\langle \psi, \psi \rangle}. \tag{2}$$

The maximum is achieved by the associated eigenvector, which by the Perron-Frobenius theorem can be taken to have positive entries [14].

From the identity $\|P - UU^T\|_F^2 = \|P\|_F - 2\mathrm{tr}(U^t P U) + k$, it can be seen that (1) is equivalent to finding a partition of $[n]$, written $[n] = \amalg_{i=1}^k \Omega_i$, which solves

$$\max_{\substack{[n] = \amalg_{i=1}^k \Omega_i \\ \mathrm{supp}(\psi_i) \subset \Omega_i}} \sum_{i=1}^k \frac{\langle \psi_i, P\psi_i \rangle}{\langle \psi_i, \psi_i \rangle}. \tag{3}$$

In terms of the Perron-Dirichlet eigenvalues, $\lambda_\Omega$, in (2), Equation (3) can thus be rewritten

$$\max_{[n] = \amalg_{i=1}^k \Omega_i} \sum_{i=1}^k \lambda_{\Omega_i}. \tag{4}$$

**Theorem 2.1.** *Let $P$ be a nonnegative, symmetric, irreducible matrix and $\Psi^* = [\psi_1 | \cdots | \psi_k]$ be the matrix whose columns are the eigenvectors corresponding to the optimal partition in (4). Then $\Psi^*$ attains the minimum in* (1).

Intuition for (4) can be gained by considering a system with $k$ species of random walkers. Each species begins in a component of the disjoint partition $[n] = \amalg_{i=1}^k \Omega_i$, and walks independently according to the transition matrix, $P$. The following lemma shows that $\lambda_\Omega$ can be interpreted as a measure of closedness for the subsystem associated to the species beginning in $\Omega$.

**Lemma 2.2.** *Let $P$ be a symmetric transition matrix and $\Omega \subset [n]$. Suppose the submatrix $P_\Omega$ is irreducible. Let $\Phi_\Omega := \frac{1}{|\Omega|}\sum_{i,j\in\Omega} P_{ij}$ be the conditional probability of staying in $\Omega$ given that the random walk starts in $\Omega$. Let $\Xi_\Omega := \max_{k\in\Omega}\sum_{j\in\Omega} P_{kj}$ be the maximal conditional probability of a walker ending in $\Omega$ given that it started at a particular state in $\Omega$. Then $\Phi_\Omega \leq \lambda_\Omega \leq \Xi_\Omega$.*

Lemma 2.2 can be proven using standard techniques from Perron-Frobenius theory. Lemma 2.2 shows that (4) can be interpreted as finding the disjoint partition of $[n]$ which are optimally closed for the $k$ species of walkers. The equivalence in Thm 2.1 implies that the block structure given in $UU^t$ also indicates these subsystems. In §4, we return to this probabilistic intuition and consider a system of $k$ species which walk independently according to the transition matrix, $P$, but an antagonistic interaction term forces disjoint support.

The eigenfunction $\psi_\Omega$ corresponding to $\lambda_\Omega$ is an approximation to the stationary distribution on the subgraph induced by $\Omega$. Thus, we interpret $\psi_\Omega$ to be a ranking of the states in $\Omega$, as in PageRank [9]. A large ranking indicates that the state is central to that subgraph. In this case, (1) (or equivalently (4)) simultaneously clusters the items in a dataset and ranks their representation within the cluster. The state for which $\psi_\Omega$ is maximal is the most representative state for the label $\Omega$.

## 3 Relaxation, a rearrangement algorithm, and numerical experiments

In this section, we describe an algorithm for approximately solving (4), based on methods in [13]. We begin by defining an appropriate relaxation. For $\alpha > 0$, and $\phi : [n] \to [0,1]$, let $\lambda^\alpha(\phi)$ be the largest eigenvalue of the perturbed matrix $P - \alpha(1-\phi)$. Observe that this eigenvalue can be written

$$\lambda^\alpha(\phi) = \max_{\psi\neq 0} \frac{\langle P\psi, \psi\rangle - \alpha\langle 1-\phi, \psi\rangle}{\langle \psi, \psi\rangle}. \tag{5}$$

For $\Omega \subset [n]$, $\lim_{\alpha\to\infty} \lambda^\alpha(\chi_\Omega) = \lambda_\Omega$ and $\lim_{\alpha\to\infty} \psi^\alpha(\chi_\Omega) = \psi(\Omega)$. Intuitively, a large value of $\alpha$ enforces localization of $\psi$ on supp$(\phi)$.

Define the relaxed eigenvalue optimization problem,

$$\Lambda_k^{\alpha,*} := \max_{\{\phi_i\}_{i=1}^k \in \mathcal{A}_k} \sum_{i=1}^k \lambda^\alpha(\phi_i) \ \text{ where } \mathcal{A}_k := \{\{\phi_i\}_{i=1}^k \colon \phi_i \colon [n] \to [0,1] \text{ and } \sum_{i=1}^k \phi_i = 1\}. \tag{6}$$

**Theorem 3.1.** *Let $k \in \mathbb{Z}^+$ and $\alpha > 0$ be fixed. Every (local) maximizer of $\Lambda_k^\alpha$ over $\mathcal{A}_k$ is a collection of indicator functions.*

Equation (6) was motivated by a relaxation for a geometric partitioning problem [15, 16, 17]. Following [13], we propose a rearrangement algorithm for the solution of (6) (Algorithm 1), which enjoys the following convergence and local optimality guarantee.

**Theorem 3.2.** *Let $\alpha > 0$. For any initialization, the rearrangement algorithm 1 terminates in a finite number of steps at a local maximum of $\Lambda_k^\alpha$, as defined in* (6)*.*

Algorithm 1 is an attractive alternative for solving the NMF objective (1) due to its simplicity and convergence guarantee; the only requirement on $P$ is that $P + P^t$ be nonnegative and irreducible. Algorithm 1 is based on geometrically rearranging the partitions based on the sub-levelsets of the relaxed eigenfunctions. Moreover, Algorithm 1 admits a natural semi-supervised extension. If a subset of the labels are known, one can fix the label membership throughout the algorithm and the convergence guarantees still remain valid. We can interpret this as a set of stationary walkers of fixed species which act as a barrier to other species. Same-species walkers will congregate around these fixed-label points.

### 3.1 Numerical experiment: MNIST handwritten digit dataset

We consider an MNIST dataset consisting of 70,000 $28 \times 28$ greyscale images of handwritten digits 0 to 9. We randomly sampled 3% of the data and used the semi-supervised variant of our algorithm described above. The remaining initialized labels were assigned randomly.

**Algorithm 1** A rearrangement algorithm for (6).

```
**Input:** An initial $\{\phi_i\}_{i=1}^k \in \mathcal{A}_k$.
**while** not converged, **do**
    For $i = 1, \ldots, k$, compute the (positive and normalized) eigenfunction $\psi_i$ corresponding to $\lambda^\alpha(\phi_i)$ in (2).
    Assign each state $v \in [n]$ the label $i = \arg\max_j \ \psi_j(v)$.
    Let $\{\phi_i\}_{i=1}^k$ be the indicator functions for the labels.
**end while**
```

For ten different random initializations, we run the algorithm until convergence and choose the lowest energy partition. In each case, the algorithm converges in approximately 20 iterations. The purity obtained, as defined in [8], is 0.961 which is comparable to the performance of state-of-the-art clustering algorithms. We note that the partitions identified for other initial configurations had similar energy and purity values. More details, including figures, can be found in [13].

## 4 Discussion and a connection to a reaction-diffusion system

In §2 we showed the equivalence of a clustering NMF of the transition matrix, $P$, as in (1), to an eigenvalue optimization problem involving Perron-Dirichlet eigenvalues of $P$, as in (4). Lemma 2.2 provides an interpretation of (4) in terms of finding subsystems for which random walkers are optimally closed. In §3, we proceeded to find a relaxation of (4), and a rearrangement algorithm for finding approximate solutions.

Here, we return to the interpretation of (4) in terms of random walkers and introduce an "antagonistic" interaction which discourages distinct species from occupying the same state. Following [18, 19], we consider an *interacting system* of $k$ species of random walkers, each of which walks independently according to $P$, and interacts according to the following rule: if two walkers of different species meet, they annihilate one another. When a walker of a given species is annihilated, another walker of the same species is chosen at random and duplicated, conserving the total number of each species.

This interacting system is modeled by the nonlinear, nonlocal, reaction-diffusion equations,

$$\frac{d}{dt}p_i = -(\Delta + \kappa V_i)p_i + \frac{\kappa}{n}\langle p_i, V_i\rangle 1, \qquad i = 1, \ldots, k, \tag{7}$$

where $\Delta = \mathrm{Id} - P$ is the Laplacian, $V_i = \sum_{j \neq i} p_j^2$ is a nonlinear potential, and $\kappa > 0$ is an interaction parameter. (7) is the $\ell^2$ gradient flow of the energy $E[p] = \frac{1}{2}\sum_i \langle p_i, \Delta p_i\rangle + \frac{\kappa}{4}\sum_{i \neq j}\langle p_i^2, p_j^2\rangle$ subject to the constraints that $\langle p_i, 1\rangle = 1$ for all $i \in [k]$. Due to the invariance of the positive orthant, if the system is initiated with each $p_i$ on the probability simplex, it will remain there for all time. Thus, $p_i(t) \in \mathbb{R}^n$ is interpreted as the probability of finding a walker of species $i$ on a given state at time $t$. The linear "diffusion" term in (7) can be interpreted as the random walk. The nonlinear potential penalizes the overlap between $p_i$ and $p_j$ for $i \neq j$. The nonlinear terms in (7) represent the interaction between different species. Since the mass of each $p_i$ is conserved, this term forces each species to dominate a subset of the states.

We conjecture that as $\kappa \to \infty$, the stationary states of (7) are equivalent to the Perron-Dirichlet eigenfunctions, attaining (2). If true, this exposes a new avenue for NMF algorithm development. In particular, if stationary states of (7) can be efficiently found, then to each state we assign the class label corresponding to the species dominating there. Many open questions remain concerning the system (7) and, in particular, its relationship to (2) and (5), but we hope that the geometric interpretation provided here could already lead to an improved understanding of other algorithms.

### Acknowledgments

Braxton Osting is supported in part by NSF grant DMS-1103959 and AFOSR MURI grant FA9550-10-1-0569. Chris White was supported by in part by AFOSR MURI grant FA9550-10-1-0569, ONR grant N000141210838, and ONR grant N000141210040.

## References

[1] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*, pages 50–57. ACM, 1999.

[2] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[3] L. Lee and D. Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13:556–562, 2001.

[4] V. Bittorf, B. Recht, C. Re, and J. A. Tropp. Factoring nonnegative matrices with linear programs. *preprint arXiv:1206.1270*, 2012.

[5] C. H. Q. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM*, volume 5, pages 606–610, 2005.

[6] R. Arora, M. Gupta, , A. Kapila, and M. Fazel. Clustering by left-stochastic matrix factorization. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 761–768, 2011.

[7] D. Kuang, C. H. Q. Ding, and H. Park. Symmetric nonnegative matrix factorization for graph clustering. In *SDM*, volume 12, pages 106–117, 2012.

[8] Z. Yang, T. Hao, O. Dikmen, X. Chen, and E. Oja. Clustering by nonnegative matrix factorization using graph random walk. In *NIPS*, 2012.

[9] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1):107–117, 1998.

[10] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

[11] E. Elhamifar, G. Sapiro, and R. Vidal. Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery. In *Advances in Neural Information Processing Systems*, pages 19–27, 2012.

[12] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[13] B. Osting, C. D. White, and É. Oudet. Minimal Dirichlet energy partitions for graphs. preprint, arxiv: 1308.4915, 2013.

[14] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[15] D. Bucur, G. Butazzo, and A. Henrot. Existence results for some optimal partition problems. *Adv. Math. Sci. Appl.*, 8:571–579, 1998.

[16] L. A. Cafferelli and F. H. Lin. An optimal partition problem for eigenvalues. *J. Sci. Comp.*, 31, 2007.

[17] B. Bourdin, D. Bucur, and É. Oudet. Optimal partitions for eigenvalues. *SIAM Journal on Scientific Computing*, 31(6):4100–4114, 2010.

[18] O. Cybulski, V. Babin, and R. Holyst. Minimization of the Renyi entropy production in the space-partitioning process. *Physical Review E*, 71(4):046130, 2005.

[19] O. Cybulski and R. Holyst. Three-dimensional space partition based on the first Laplacian eigenvalues in cells. *Physical Review E*, 77(5):056101, 2008.